Universitat Oberta de Catalunya

Analysis and Application of clustering and visualization methods of computed tomography radiomic features to contribute to the characterization of patients with non-metastatic Non-small-cell lung cancer.

Maria Mercedes Serra

Área 2, Subárea 11
Máster en Bioinformática y Bioestadística

Daniel Fernández
(Carles Ventura Royo)

02 de junio de 2022

NO LICENSE

FINAL WORK CARD

Title:	Analysis and Application of clustering and visualization methods of computed tomography radiomic features to contribute to the characterization of patients with non-metastatic Non-small-cell lung cancer.
Author:	Maria Mercedes Serra
Tutor:	Daniel Fernández
SRP:	Carles Ventura Royo
Date of delivery:	02 de junio de 2022
Studies:	Máster en Bioinformática y Bioestadística
Area:	Área 2, Subárea 11
Language:	Ingles
Number of credits:	15
Keywords:	Radiomics, clustering, data visualization

Abstract

Abstract

Background: The lung is the most common site for cancer and has the highest worldwide cancer-related mortality. Routine study of patients with lung cancer usually includes at least one computed tomography (CT) study previous to the histopathological diagnosis. In the last decade the development of tools that help extract quantitative measures from medical imaging, known as radiomic characteristics, have become increasingly relevant in this domain, including mathematically extracted measures of volume, shape, texture analysis, etc. Radiomics can quantify tumor phenotypic characteristics non-invasively and could potentially contribute with objective elements to support these patients' diagnosis, management and prognosis in routine clinical practice.

Methodology: LUNG1 dataset frommUniversity of Maastricht and publicly available in The Cancer Imaging Archive was obtained. Radiomic feature extraction was performed with pyRadiomics package v3.0.1 using CT scans from 422 non-small cell lung cancer (NSCLC) patients, including manual segmentations of the gross tumor volume. A single data frame was constructed including clinical data, radiomic features output, CT manufacturer and study date acquisition information. Exploratory data analysis, curation, feature selection, modeling and visualization was performed using R Software. Model based clustering was performed using *VarselLCM* library both with and without wrapper feature selection.

Results: During exploratory data analysis lack of independence was found between histology and age and overall stage, and between survival curves and scanner manufacturer model. Features related to the manufacturer model were excluded from further analysis. Additional feature filtering was performed using the MRMR algorithm. When performing clustering analysis both models, with and without variable selection, showed significant association between partitions generated and survival curves, significance of this association was greater for the model with wrapper variable selection which selected only radiomic variables. original_shape_VoxelVolume feature showed the highest discriminative power for both models along with log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis. Clusters with significant lower median survival were also related to higher Clinical T stages, greater mean values of original_shape_VoxelVolume, log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis and lower mean wavelet.HHl_glcm_ClusterPro-minence. A weaker relationship was found between histology and selected clusters.

Conclusions: Potential sources of bias given by relationship between different variables of interest and technical sources should be taken into account when analyzing this data set. Aside from original_shape_VoxelVolume feature, texture features applied to images with LoG and wavelet filters where found most significantly associated with different clinical characteristics in the present analysis.

Contents

List of Figures

List of Tables

Chapter 1

Abstract

Background: The lung is the most common site for cancer and has the highest worldwide cancer-related mortality. Routine study of patients with lung cancer usually includes at least one computed tomography (CT) study previous to the histopathological diagnosis. In the last decade the development of tools that help extract quantitative measures from medical imaging, known as radiomic characteristics, have become increasingly relevant in this domain, including mathematically extracted measures of volume, shape, texture analysis, etc. Radiomics can quantify tumor phenotypic characteristics non-invasively and could potentially contribute with objective elements to support these patients' diagnosis, management and prognosis in routine clinical practice.

Methodology: LUNG1 dataset frommUniversity of Maastricht and publicly available in The Cancer Imaging Archive was obtained. Radiomic feature extraction was performed with pyRadiomics package v3.0.1 using CT scans from 422 non-small cell lung cancer (NSCLC) patients, including manual segmentations of the gross tumor volume. A single data frame was constructed including clinical data, radiomic features output, CT manufacturer and study date acquisition information. Exploratory data analysis, curation, feature selection, modeling and visualization was performed using R Software. Model based clustering was performed using *VarselLCM* library both with and without wrapper feature selection.

Results: During exploratory data analysis lack of independence was found between histology and age and overall stage, and between survival curves and scanner manufacturer model. Features related to the manufacturer model were excluded from further analysis. Additional feature filtering was performed using the MRMR algorithm. When performing clustering analysis both models, with and without variable selection, showed significant association between partitions generated and survival curves, significance of this association was greater for the model with wrapper variable selection which selected only radiomic variables. original_shape_VoxelVolume feature showed the highest discriminative power for both models along with log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis. Clusters with significant lower median survival were also related to higher Clinical T stages, greater mean values of original_shape_VoxelVolume, log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis and lower mean wavelet.HHl_glcm_ClusterPro-

minence. A weaker relationship was found between histology and selected clusters.

Conclusions: Potential sources of bias given by relationship between different variables of interest and technical sources should be taken into account when analyzing this data set. Aside from original_shape_VoxelVolume feature, texture features applied to images with LoG and wavelet filters where found most significantly associated with different clinical characteristics in the present analysis.

Value: This work highlights the relevance of analyzing clinical data and technical sources when performing radiomic analysis. It also goes through the different steps needed to extract, analyze and visualize a high dimensional dataset of radiomic features and describes associations between radiomic features and clinical variables establishing the base for future work.

Chapter 2

Introduction

2.1 Background and Rationale

The lung is the most common site for cancer and has the highest worldwide cancer-related mortality (Jani et al., 2021). Non-small-cell lung cancer (NSCLC), a heterogeneous class of tumors, represents approximately 85% of all new lung cancer diagnoses (Gridelli et al., 2015). Routine study of patients with lung cancer usually includes at least one computed tomography (CT) study previous to the histopathological diagnosis. In clinical practice, the report of these studies includes the description of qualitative characteristics appreciated by the radiologists, sometimes including in addition some manually measured diameters to describe the most prominent lesions (Purandare et al., 2015).

In the last decade the development of tools that help extract quantitative measures from medical imaging, known as radiomic characteristics, have become increasingly relevant in this domain. These include mathematically extracted measures of volume, shape, texture analysis, etc. Radiomics characteristics extracted from non-invasive medical imaging studies could contribute with objective elements to support these patients' diagnosis, management and prognosis in routine clinical practice (Scrivener et al., 2016). Even though relevant biomarkers related to radiomics characteristics have been previously identified, the amount of variables that may be extracted from a CT study to characterize a lesion are practically infinite and there is still much room to explore in this domain (Aerts et al., 2019; Junior et al., 2018; Luna et al., 2022). There is a known complexity associated to the robust extraction of radiomics characteristics from medical imaging itself given the variability that exists regarding image equipment and parameters of acquisition, lesion segmentation techniques, post-processing pipelines and software for radiomic feature extraction itself (Scrivener et al., 2016; Rizzo et al., 2018). In response to this, in the last couple of years groups of experts in the field joined, aiming to define standard pipelines to proceed in order to achieve this quantitative characterization of medical imaging in the most robust way possible across different studies (Zwanenburg et al 2020). There are still a variety of accepted approaches to achieve radiomic feature extraction, so the most relevant issue is still the detailed documentation of every step followed to get these variables (van Timmeren et al., 2020).

Once radiomics features are extracted, given the high dimensionality and variability of this

data, the analysis and use of an optimal approach for variable selection, dimensionality reduction and clustering techniques become paramount to understand discover the underlying relevant information (Lee et al., 2019; Liu et al., 2020; Yousefi et al., 2019).

Cluster analysis is a generic term used for procedures which seek to uncover or discover groups in data (Landau et al., 2011). When dealing with multivariate data, It is recommended to include a pre-processing phase in which the irrelevant features are removed in order to enhance the quality of clustering. Redundant variables are usually correlated with other variables. Different techniques are usually used with the aim of selecting main relevant features; the main distinction resides in whether the selection process is performed independently or jointly with the learning/modeling procedure. (Fop et al., 2018; Fournier et al., 2021). Most common clustering methods include probabilistic and generative models (mixture model-based clustering), distance-based methods (e.g. hierarchical, k-means), and Density and Grid-Based Methods. Probabilistic and generative models assume the data has an underlying mixture of models explaining its distribution and try to find the parameters that best describe these models. Main advantage of model-based clustering methods is that they help us not only group data but understand the underlying model explaining the data and accounts for uncertainty in group assignments. (Aggarwal et al., 2014)

2.2 Objetives

2.2.1 Main Objectives

- Identify groups that contribute to the characterization of NSCLC using clinical data and radiomics characteristics extracted from baseline CT images.
- Identify elements that contribute to the prognosis of patients with NSCLC using clinical data and radiomics characteristics extracted from baseline CT images.

2.2.2 Specific Objectives

- Apply recommended techniques and documentation for reproducible radiomic feature extraction from CT images.
- Perform an exploratory data analysis and data curation including identification and imputation of missing values and outliers.
- Identify sources of bias within the data set and select the most robust features.
- Analyze different methods and available libraries for multivariate feature selection, select a method and apply it for the analysis.
- Analyze different methods and available libraries for model based clustering, select a method and apply it for the analysis.
- Apply different visualization methods for every step of the analysis.

2.3 Approach and Methodology

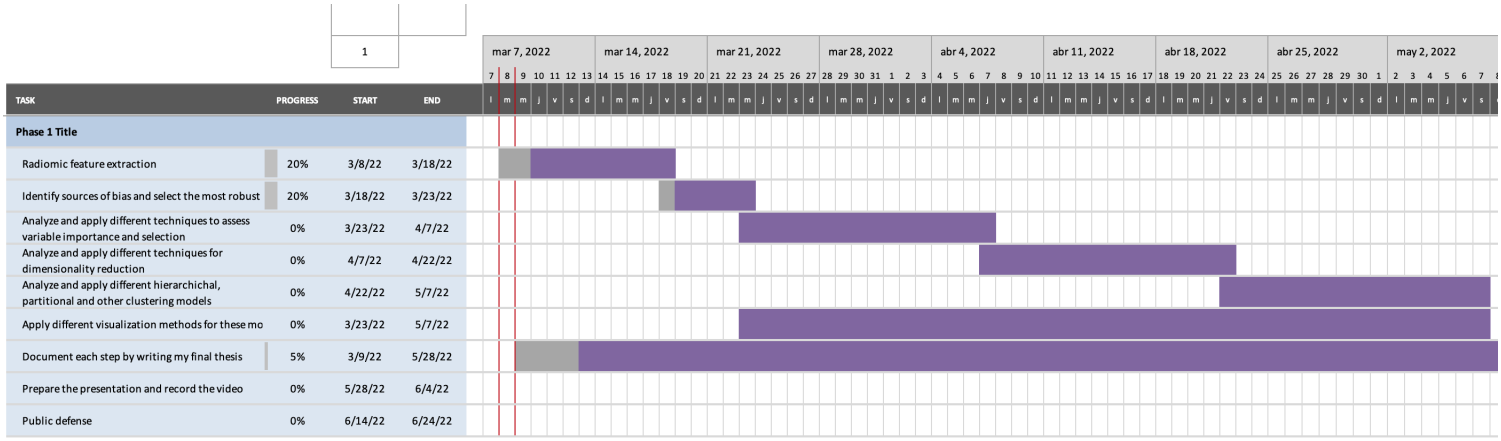
The data collection used for this work was provided by the University of Maastricht and is documented and publicly available in The Cancer Imaging Archive (Aerts et al., 2019; Aerts et al., 2014; Clark K et al., 2013). It contains pre-treatment CT scans from 422 non-small cell lung cancer (NSCLC) patients, including manual segmentations of the gross tumor volume done by a radiation oncologist. Associated clinical data is also available.

For image format conversion, from DICOM to NIFTI I used plastimatch extension in 3D Slicer (Sharp et al., 2010). Once files were converted I used the main CT image and GTV-1 mask, corresponding to the primary gross tumor volume (Aerts et al., 2019.) as input to extract the radiomic features corresponding to the primary lesion. For radiomic feature extraction I used pyRadiomics package v3.0.1 (van Griethuysen et al., 2017), currently one of the most commonly used packages for radiomics analyses (van Timmeren et al., 2020).

With clinical and radiomic data frames available, data analysis and modeling was then done using R Software (R Core Team, 2020). Steps included in the analysis were: exploratory data analysis, feature selection, standardization, dimensionality reduction and clustering analysis.

2.4 Planning, Milestones and Calendar

1. Apply recommended techniques for reproducible radiomic feature extraction from CT images using software available.
2. Perform and exploratory data analysis and curation. Use R software, explore specific packages.
3. Identify sources of bias and select the most robust features.
4. Analyze different methods and available libraries for multivariate feature selection, select a method and apply it for the analysis. Use R software, explore specific packages.
5. Analyze different methods and available libraries for model based clustering, select a method and apply it for the analysis. Use R software, explore specific packages.
6. Apply different visualization methods for these models. Use R software, explore specific packages.
7. Document each step by writing my final thesis
8. Prepare the presentation and record the video
9. Public defense



2.5 Summary of the Products

The product of this project will be the documentation of the methodology followed to extract radiomics features from a publicly available set of computed tomography images and segmentations from patients with non-small cell lung cancer, including technical challenges found. The documentation of exploratory data analysis and curation of this data set. The revision and documentation of methodology followed to apply different dimensionality reduction and clustering techniques, and different visualization methods for these models.

2.6 Brief Description of Following Chapters

The third chapter of this thesis will include a revision of the state of the art regarding:

- Medical context,
- Radiomics,
- Clustering techniques.

The fourth chapter will consist of the documentation of the methodology followed for the present work. It will include original data source, the steps followed in order to perform radiomic feature extraction from computed tomography images, an exploratory data analysis and curation of the data set, and the methodology used to apply clustering techniques along with the visualization of these models.

The fifth chapter will present the results obtained after data analysis.

The fifth chapter includes a discussion of the results obtained in the present work regarding the state of the art and evaluating the potential relevance.

Chapter 3

State of the Art

3.1 Medical Context

The lung is the most common site for cancer and has the highest worldwide cancer-related mortality (Jani et al., 2021). Lung cancer is broadly classified regarding two primary groups, small versus non–small cell type. Non-small-cell lung cancer (NSCLC) is a heterogeneous class of tumors that represents approximately 85% of all new lung cancer diagnoses (Gridelli et al., 2015). The main subtypes of NSCLC are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Though prior to 1990s squamous cell lung carcinoma was the most common histologic subtype, Adenocarcinoma has become more frequent in the last decades responding to smoking and other environmental changes. Though Adenocarcinoma is currently the most prevalent lung cancer histotype globally, this switch has not yet been established in every country (Barta et al., 2019). A high proportion of NSCLC are still classified as not otherwise specified (NOS) usually responding to insufficient material and/or undifferentiated characteristics (Righi et al., 2014).

The Union for International Cancer Control (UICC) regularly publishes and updates the internationally accepted standard for cancer staging. TNM staging is the internationally accepted system used to characterize the anatomic extent of disease, it helps determine an overall malignant tumor stage given the definition of its 3 individual components:

- T category (describes tumor location and size)
- N category (describes regional lymph nodes metastasis)
- M category (describes distant metastasis)

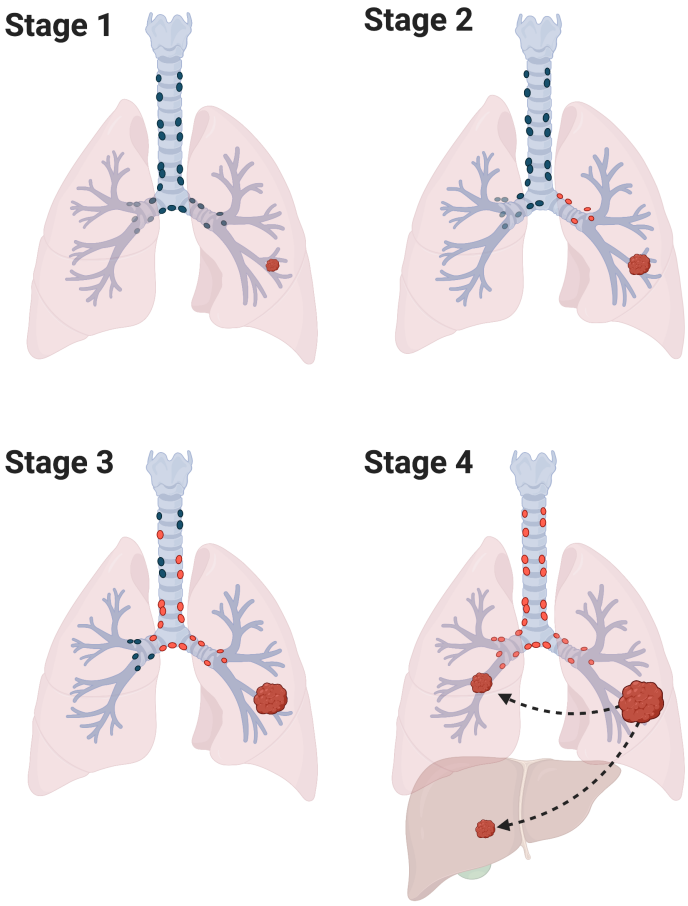


Figure 3.1: Figure representing main lung cancer stage groups. (Figure use authorized for academic purpose).

TNM staging is usually evaluated clinically, mainly supported by complementary imaging studies cTNM, and may then be followed by a pathologic assessment (pTNM). Among others, the goal of this staging system is to ease communication between experts, ease identification of best standard of care and help clinical experts in treatment decision and estimating prognosis. Current edition, TNM classification UICC 8th edition, was published in 2017 (Brierley et al.,2017). In table 3.1 we can see the detailed current classification system.

T.Subcategory	M.Subcategory	N0	N1	N2	N3
T1a	M0	Stage IA1	Stage IIB	Stage IIIA	Stage IIIB
T1b	M0	Stage IA2	Stage IIB	Stage IIIA	Stage IIIB
T1c	M0	Stage IA3	Stage IIB	Stage IIIA	Stage IIIB
T2a	M0	Stage IB	Stage IIB	Stage IIIA	Stage IIIB
T2b	M0	Stage IIA	Stage IIB	Stage IIIA	Stage IIIB
T3	M0	Stage IIB	Stage IIIA	Stage IIIB	Stage IIIC
T4	M0	Stage IIIA	Stage IIIA	Stage IIIB	Stage IIIC
Any T	M1a	Stage IVA	Stage IVA	Stage IVA	Stage IVA
Any T	M1b	Stage IVA	Stage IVA	Stage IVA	Stage IVA
Any T	M1c	Stage IVB	Stage IVB	Stage IVB	Stage IVB

Table 3.1: Lung Cancer TNM classification system.

Regarding standard of care guidelines, NSCLC are broadly divided in two groups: early stage/locally advanced (non-metastatic), and metastatic. In current guidelines a pretreatment pathological diagnosis, which implies some kind of invasive procedure to obtain the sample, is recommended prior to any curative treatment in patients with clinical stages I–III lesions (Postmus et al., 2017). Routine study of patients with lung cancer usually includes at least one computed tomography (CT) study previous to the histopathological diagnosis. In clinical practice, the report of these studies includes the description of qualitative characteristics appreciated by the radiologists, sometimes including in addition some manually measured diameters to describe the most prominent lesions (Purandare et al., 2015). In the last decade the development of tools that help extract quantitative measures from medical imaging, known as radiomic characteristics, have become increasingly relevant in this domain. Radiomics can quantify tumor phenotypic characteristics non-invasively and could potentially contribute with objective elements to support these patients' diagnosis, management and prognosis in routine clinical practice (Scrivener et al., 2016). We should always take into account though, that CT studies imply a dose of ionizing radiation for the patient each time this study is indicated.

3.2 Radiomics

In the last decade the development of tools that help extract quantitative measures from medical imaging, known as radiomic characteristics, have become increasingly relevant in this domain. These include mathematically extracted measures that may help describe lesion phenotypic characteristics non-invasively and their changes over time when performed on serial imaging (Maeyerhoefer et al. 2020). Main advantage over tissue biopsies, is that while this are usually limited to the specific tumor site were the sample was obtained, radiomic features evaluate the lesion as a whole and may therefore better describe and account for lesion heterogeneity (Papanikolaou et al., 2020). In addition, this kind of measurement is easier to repeat in follow-up studies to give an objective description of disease evolution and account for subtle lesions changes

that could be hard to detect with the eye without adding invasive procedures. Radiomic features have been previously associated with diagnosis, staging, classification, response to different therapies and predicting survival (Fournier et al., 2021).

Features are calculated taking in account an image segmentation mask that identifies the voxels located within a region of interest (ROI).

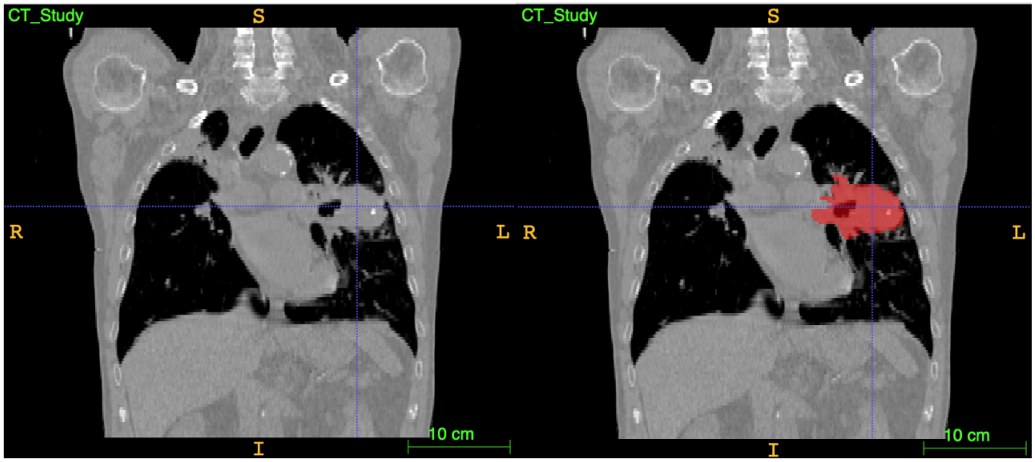


Figure 3.2: Example of image segmentation identifying the region of interest within a chest CT study of a patient with NSCLC from a dataset used in the current work.

Segmentation masks may come from previous manual, automatic or semiautomatic definition of the region of interest. Though manual segmentation by experts is still considered the gold standard, intra and inter-operator variability add potential sources of bias. Regarding feature extraction, the ROI itself consists of two masks, a morphological mask (a binary mask just defining voxels included within the ROI), an intensity mask (including intensity values for each of the voxels within the ROI), depending on the type of feature one mask and/or the other will be used (Zwanenburg et al., 2019).

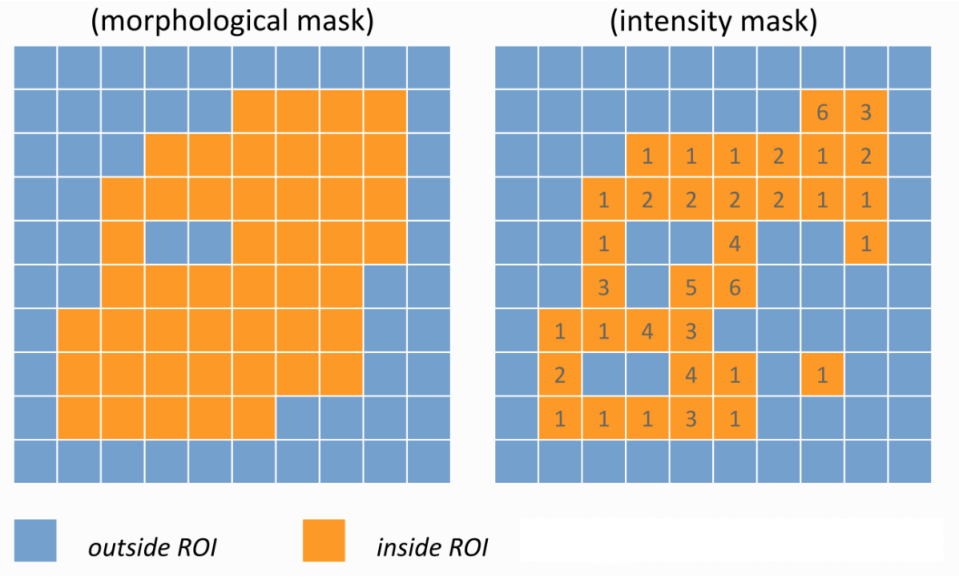


Figure 3.3: Image modified from source (Zwanenburg et al., 2019) intended for academic purpose.

Specially when working with data from different equipment and protocols, pre-processing signal normalization is necessary to bring signal intensities to a common scale, without distorting differences in the ranges of values (Papanikolaou et al., 2020). There is still a variety of accepted approaches to achieve radiomic feature extraction and many platforms available that may add variance in the results, so the most relevant issue is still the detailed documentation of every step followed to get these variables (van Timmeren et al., 2020, Fornacon-Wood et al. 2020).

Main family of radiomic features include:

- Shape-based (2D and 3D)/ Morphological features: describe geometric properties of the morphological mask.
- First Order Statistics/ Histogram features: describe the distribution of voxel intensities within the image region defined by the mask through commonly used and basic metrics.
- Texture features:
  - Gray Level Co Occurrence Matrix (GLCM): describes spatial relationships of pairs of pixels or voxels with predefined gray-level intensities, in different directions.
  - Gray Level Run Length Matrix (GLRLM): quantifies gray level runs, defined as the length of consecutive pixels that have the same gray level intensity.
  - Gray Level Size Zone Matrix (GLSZM): quantifies gray level zones, defined as a the number of connected voxels that share the same gray level intensity. Contrary to GLCM and GLRLM, the GLSZM is rotation independent, with only one matrix calculated for all directions in the ROI.
  - Gray Level Dependence Matrix (GLDM): quantifies gray level dependencies in an image defined as a the number of connected voxels within a distance that are dependent on the center voxel.

Detailed description for each of these radiomic features, and additional ones, may be found within pyradiomics documentation `https://pyradiomics.readthedocs.io/en/latest/features.html#` and within the imaging biomarkers standardization initiative (IBSI) documentation (Zwanenburg et al., 2019).

Except for the shape descriptors that are independent of gray value, and are extracted from the label mask, all other features may be calculated on the original image, as well as on images transformed using different filters such as wavelet and Laplacian of Gaussian (LoG) filters (van Griethuysen et al., 2017). Wavelets is a category of filtering methods that includes different combinations of high-pass and low-pass filters applied to the different directions. These filters help enhance intricate patterns in the data that are difficult to quantify by eye as they emphasize specific image characteristics such as edges and blobs (Depeursinge et al., 2020, Papanikolaou et al., 2020). Some feature families, as histogram or texture, require additional pre-processing steps before feature calculation including prior discretization of intensities into fixed gray level bins (Zwanenburg et al., 2019).

Once radiomic feature extraction is performed this leads usually to high dimensional datasets that may be statically analyzed aiming to search for association with meaningful clinical endpoints, then a biological association may follow. Before refining analysis or data modeling, given the high dimensionality of these datasets feature selection is crucial. In addition, radiomic features tend to be highly redundant as they derive from multiple slightly different mathematical formulas or even same measurements obtained by applying different image filters.

It is worth noting that different image formats are available for medical imaging, DICOM is (Digital Imaging and Communications in Medicine) is a standard protocol for the management and transmission of medical images and related data. This format ensures saving image information itself along with associated metadata that registers relevant patient and device, and acquisition protocol information in a standard way. Special formats to save segmentation information within DICOM standard may be used, including DICOM SEG and RTSTRUCT. One of the disadvantages of DICOM format is that it uses a single file for each slice of each series of a study. NIfTi and Nrrd (Nearly Raw Raster Data) were both designed to simplify medical image workflows specially in the context of research. The raw image data is saved as a 3d image so there is a single file for each series which makes working with images much faster. On the contrary, these formats preserve only essential metadata as image geometry, and most additional relevant metadata is handled using csv or json files. Segmentations may be saved in NIfTi and Nrrd as well. Most post-processing and radiomics feature extraction platforms take NIfTi and Nrrd format as input, so previous image format conversion is usually necessary.

3.3 Cluster Analysis

It is within human nature to try to classify individuals or objects into groups with the goal of organizing sets of data with class labels that describe similarities and differences within and between groups (Everitt et al., 2011). These classifications will be usually judged regarding the usefulness. Cluster analysis is a generic term used for procedures which seek to uncover groups in data (Landau et al., 2011). Clustering of data is usually evaluated taking in account

the homogeneity within groups, the separation between groups and most importantly, as previously mentioned, its usefulness. For the human brain it is easier to identify groups in 1, 2 or even 3 dimensions, but the problem of identifying groups becomes more difficult when facing multivariate datasets. One common application of clustering techniques themselves, is to create compact data representations which are easier to process and interpret by the human brain, closely related to dimensionality reduction techniques (Aggarwal et al., 2014).

As with other techniques, exploratory data analysis is an important initial step. First for understanding the data and relevant questions that may arise, evaluating the quality regarding missing values and outliers, and identifying incoherences. In addition, graphical displays of data can be useful for suggesting the data may in fact contain clusters and may further benefit from applying formal cluster analysis, for example when identifying some degree of multimodality (Landau et al., 2011).

When dealing with multivariate data, It is recommended to include a pre-processing phase in which the irrelevant features are removed in order to enhance the quality of clustering. Irrelevant variables can be divided in uninformative, mainly noisy variables, or redundant, variables that provide similar information to the obtained by another variable therefore not contributing to a parsimonious model. Redundant variables are usually correlated with other variables. Different techniques are usually used with the aim of selecting main relevant features; the main distinction resides in whether the selection process is performed independently or jointly with the learning/modeling procedure. The first approach corresponds to filter methods, this aim is to expose relationships between features as well as correlation to the class of interest but the selection is performed as a pre o post-processing step but independent from the statistical modeling. These techniques are usually easy to implement, computationally efficient and are usually better in preventing overfitting, but may exclude variables that seemed irrelevant on their own but were relevant contributors to the model as a whole. The second approach is the wrapper methods that perform variable selection and model training simultaneously. As each potential subset is found it is tested against the learning algorithm and scored. Though more prone to overfitting and computationally more expensive, this often provides superior performance results so they've gained popularity (Fop et al., 2018; Fournier et al., 2021).

Dimensionality reduction techniques may also be used as a pre-processing step or directly into a clustering algorithm in order to enhance the quality of the analysis or gain additional insights (Aggarwal et al., 2014).

Though in general terms, variable standardization is usually recommended to avoid bias coming from different variable units and scales, this is not necessarily mandatory and can sometimes be misleading as weights may be reduced for variables that contributed to cluster isolation. Analyzing within cluster variation to decide this may be useful but this is not always possible as groups are not always known in advance (Everitt et al., 2011; Haga et al, 2019).

Most common clustering methods include probabilistic and generative models (mixture model-based clustering), distance-based methods (e.g. hierarchical, k-means), and Density and Grid-Based Methods (Aggarwal et al., 2014). Different to heuristic methods, probabilistic and generative models assume the data has an underlying model explaining its distribution (e.g. mixture of gaussians, bernoulli, poisson, etc) and the parameters of these models are then calculated using the Expectation Maximization algorithm. Main advantage of model-based

clustering techniques is that they help us understand the underlying data model explaining the clusters and it also accounts for uncertainty when assigning an observation to a specific cluster. Different to kmeans where distance to cluster centroid is the main condition to defining observation assignment, model-based methods accounts for each cluster maybe having different size, variance and direction.

Chapter 4

Methodology

4.1 Original Data

The data collection used for this work was provided by the University of Maastrich and is documented and publicly available in The Cancer Imaging Archive (Aerts et al., 2019; Aerts et al., 2014; Clark K et al., 2013). This data collection contains pre-treatment CT scans from 422 non-small cell lung cancer (NSCLC) patients, including manual segmentations of the gross tumor volume done by a radiation oncologist. Associated clinical data is also available, the clinical data used in the present work is the revised version from 2019, csv entitled "NSCLC Radiomics Lung1.clinical-version3-Oct 2019".

When downloading image files from NCIA a metada.csv file is also included. This file describes the main DICOM metadata associated with the files downloaded. As documented by Aerts et al., 2019 RTSTRUCT and SEG files include the manual segmentations of the gross tumor volume and selected anatomical structures (i.e., lung, heart and esophagus) done by a single radiation oncologist. DICOM SEG objects contain only a subset of annotations available in RTSTRUCT.

CT images, segmentations, DICOM metadata and clinical data to reproduce this work may be found in `https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics`.

4.2 Radiomics Features Extraction

For radiomic feature extraction I used pyRadiomics package v3.0.1 (van Griethuysen et al., 2017), currently one of the most commonly used packages for radiomics analyses (van Timmeren et al., 2020). Aside from calculating features, the pyradiomics package includes pre-processing steps necessary for a robust radiomic feature extraction. The output information on used image and mask, as well as applied settings and filters are included in the output thereby enabling fully reproducible feature extraction (Van Griethuysen et al., 2017).

In order to be able to analyze images with pyRadiomics where DICOM format is not allowed as input, I converted CT studies and their corresponding RTSTRUCT segmentations to nrrd format using plastimatch extension in 3D Slicer (Sharp et al., 2010). Once files were converted I

used the main CT image and GTV-1 mask, corresponding to the primary gross tumor volume (Aerts et al., 2019.) as input to extract the radiomic features corresponding to the primary lesion. I run PyRadiomics using batch-extraction mode from the command line. For image pre-processing specifications I used recommended default parameters available on pyRadiomics GitLab

`https://github.com/AIM-Harvard/pyradiomics/blob/master/examples/exampleSettings/exampleCT.yaml`, including all default features of the different available classes. Normalization was applied as a pre-processing step as well, as the dataset included images from different scanners with slightly varying protocols.

Software.Parameters	Value
Versions_PyRadiomics	v3.0.1.post13+g2e0b76e
Versions_Numpy	1.21.5
Versions_SimpleITK	2.1.1
Versions_PyWavelet	1.2.0
Versions_Python	3.7.12
Configuration_EnabledImageTypes	{'Original': {}, 'LoG': {'sigma': [1.0, 2.0, 3.0, 4.0, 5.0]}, 'Wavelet': {}}
minimumROIDimensions	2
minimumROISize	None
normalize	True
normalizeScale	500
removeOutliers	None
resampledPixelSpacing	[1, 1, 1]
interpolator	sitkBSpline
preCrop	False
padDistance	10
distances	[1]
force2D	False
resegmentRange	None
binWidth	25
voxelArrayShift	1000

Table 4.1: Table describing software version and parameters used for radiomic features extraction.

Once feature extraction was complete, a single data frame with one row per subject was constructed including clinical data and radiomic features output. In addition, specific metadata regarding CT manufacturer and study date acquisition information was subsetted from metadata.csv file available from NCIA download. This main data frame was used for further analysis. For full reproducibility of this work this data frame is available in `https://github.com/mechyserra/TFM_MSERRA_2022/tree/main/data`.

4.3 Data Analysis

Data analysis and modeling was performed using R Software (R Core Team, 2020). Different R libraries used to work on different steps of data analysis are detailed on the following table. For full reproducibility of this work all code used to perform the analysis is available in `https://github.com/mechyserra/TFM_MSERRA_2022`.

Task	R.libraries.used
Data wrangling	base-R, tidyverse, lubridate
Data visualization	tidyverse, naniar, DataExplorer, plotly, patchwork
Uni and bivariate summary statistics	base-R, skimr, dlookr, crosstable
Tables styling	kableExtra, flextable
Missing values imputation	missForest
Outliers detection	HDoutliers
Feature filtering	caret, praznik, mRMRe
Data transformation	base-R
Clustering and evaluation	VarSelLCM cluster
Correspondence analisis	ca, factoextra
Survival analisis	survival, survminer

Table 4.2: Table describing main R libraries used for data analysis divided by tasks.

4.3.1 Exploratory Data Analysis

As an initial step, an exploratory data analysis was performed. Main research questions defined were whether radiomic variables are related with histology class or not, whether radiomic variables are related to survival or not. Exploratory analysis was performed taking these two research questions into account. The dimensions of the data frame and data types were described using R-base functions (R Core Team, 2020). I evaluated the presence of missing values for the whole data frame, number of missing values and completeness rate for variables with at least one missing value with *skim* functions from *Skimr* library (Elin Waring et al., 2022). Barplots displaying univariate and multivariate joint missing values were generated using *gg_miss_upset* function from *naniar* library (Nicholas Tierney et al., 2021).

The univariate distribution of different categorical variables was described using absolute and relative frequency tables for factor variables, generated with *univar_category* function from *dlookr* library (Choonghyun Ryu, 2022), and univariate barplots representations using *ggplot* functions from *tidyverse* library (Wickham et al., 2019). The univariate distribution of continuous variables was described using mean, standard deviation, standard error of the mean and interquartile range using *describe* function from *dlookr* library; skewness and kurtosis measurements were evaluated as well to asses asymmetry (Choonghyun Ryu, 2022) and Shapiro-Wilk test to test for normality from R-base functions (R Core Team, 2020). Histograms and qqplots were used to

ease visualization of all this information using *ggplot* functions from *tidyverse* library (Wickham et al., 2019).

Bivariate distributions were represented using boxplots, stacked or dodge barplots using *ggplot* functions from *tidyverse* library (Wickham et al., 2019) and plot functions form *DataExplorer* library (Boxuan Cui, 2020). Independence of different variables was tested using Fisher's exact or Chi-squared test as appropriate between categorical variables; (Welsch) Two sample t-test or Wilcoxon rank sum exact test as appropriate between numerical and dichotomous categorical variables; One-way analysis of variance and Kruskal-Wallis rank sum test as appropriate between numerical and polytomous categorical variables. Tests were performed using R-base functions (R Core Team, 2020) and *crosstable* function and library (Dan Chaltiel, 2022).

Spearman and Pearson correlation coefficients were used to evaluate correlation between ordinal and continuous variables, respectively. Heatmaps were used to represent multivariate correlation coefficients using *plot_correlation* function from *DataExplorer* library (Boxuan Cui, 2020).

To evaluate the relationship between survival and different categorical variables, Kaplan Meyer curves were estimated using Survival.time and deadstatus.event variables. Survival curves were compared using log-rank test. To evaluate relationships with different numerical variables Cox proportional hazards models were fitted and significance was evaluated using Wald test. All survival analysis was performed using *Survival* and *survminer* packages (Therneau T, 2022; Kassambara A et al., 2021).

4.3.2 Missing values imputation

Subjects missing values on target variables were excluded. Rest of missing data was treated using *missForest* imputation, a non parametric iterative imputation method that uses random forest algorithm to replace missing values, applicable to various variable types (Stekhoven et al., 2012). It initializes replacing missing values with mean and mode, for continuous and categorical variables respectively, then it sorts the variables according to the amount of missing values and starts with the variable with the lowest amount as the response variable. For each defined response variable, it predicts missing values with a random forest algorithm trained using observed parts of the dataset and assigns a new imputation value. Then, it calculates the error between prior imputation and new imputation and iterates till the difference between previous imputed data matrix and current minimizes (it stops when the difference increases and keeps the last iteration values except when it stops due to max iterations defined). The normalized root mean squared error (NRMSE) is used to evaluate performance regarding continuous variables:

$$NRMSE = \sqrt{(\frac{mean((X_{true} - X_{imp}))^2}{var(X_{true})})} \tag{4.1}$$

Where $Xtrue$ is the complete data matrix, and $Ximp$ the imputed data matrix.

The proportion of falsely classified (PFC), over the categorical missing values, is computed to evaluate performance regarding categorical variables. As we generally do not know Xtrue, *missForest* provides an estimate of the imputation error based on Out-of-Bag (OOB) error estimate of random forest (Stekhoven et al., 2012).

4.3.3 Outliers detection and imputation

After performing missing values imputation multivariate outlier detection was performed using *HDoutliers* a library that implements Leland Wilkinson's hdoutliers Algorithm for Outlier Detection (Chris Fraley, 2022; Wilkinson, L. 2017), a distributional model that uses probabilities to determine outliers, specifically designed for use in multidimensional data with both continuous and categorical variables, including non-Normal distributions. As a first step data is transformed, each categorical variable is dummy coded with the amount of columns given by the total categories on the variable, then multiple correspondence analysis is calculated separately for each variable and their first component is saved to replace the categorical variable with this continuous transformation. If the number of variables is greater than 10000, random projections are used to reduce the number of columns. Then normalization is applied to all the variables of the resulting matrix. As a second step, if the number of observations exceeds the maximum number of rows (default = 10000), the data is divided into a list of exemplars (observations representing a neighborhood) and an associated list of members within the neighborhood radius of each exemplar. This aims to reduce the number of nearest-neighbor computations in high dimensional datasets. When observations do not exceed the maximum number of rows, then the result is a list where each observation is an exemplar with no additional members in the neighborhood. As a last step, nearest-neighbor distances are calculated between all pairs of exemplars and exponential distribution is fitted to the upper tail of these distances. The $1 - \alpha$ point of the fitted cumulative distribution function is calculated and members of exemplars that fall further from this cutoff, are flagged as potential outliers.

Within *HDoutliers* a library, these steps can be individually performed using *dataTrans*, *getHDmembers* and *getHDoutliers* functions or all together using *HDoutliers* function directly (Chris Fraley, 2022).

4.3.4 Radiomic features selection and transformation

In order to perform radiomic feature selection, as a first step I filtered features related to the manufacturer model, using Wilcoxon rank sum exact test and α value of 0.05. I decided not to correct α for multiple comparisons in this case, in order to favor exclusion of any possibly related features aiming to reduce bias from this source. As a second step I searched for near cero variance features using *nearZeroVar* function from the*caret* library (Max Kuhn, 2022). Then I used the maximum relevance minimum redundancy algorithm aiming to preserve features highly correlated to the main target variables of interest while eliminating features highly correlated between them (Peng et al., 2005). In order to do this I used *MRMR* function from *praznik* library (Miron B. Kursa, 2021) to select features relevant for Histology class, and *survival* and *mRMRe* libraries to select features relevant for survival analysis (Therneau T, 2022; N De Jay

et al., 2012). Different libraries were used for these steps as *praznik* does not allow working with survival objects and *mRMRe* on the other hand does not perform MRMR for multi-class categorical variables. Initially 20 features were selected for each target. Duplicates were removed for features selected in both cases. Variables still highly correlated within this subset were further filtered using textitfindCorrelation function from the textitcaret library with a threshold of 0.95 as measured by Pearson correlation coefficient (Max Kuhn, 2022).

Given their mostly right-skewed distribution, a log transformation was applied to all selected radiomic features after adding a constant value related to the minimum negative found within these variables.

4.3.5 Model-based Clustering

Age, gender and selected radiomic features were used to perform model-based clustering, rest of clinical and scanner variables were evaluated against generated partitions. To avoid unwanted bias weights coming from different scales, an additional min-max scaling was done including all numeric variables (selected radiomics and age) before adjusting the clustering model.

Model-based clustering was performed using *VarSelCluster* function from *VarselLCM* library (Marbac M et al., 2019). This algorithm was applied searching for the best model in the range of 2 to 10 clusters, and both with and without wrapper variable selection. BIC criterion was used to select the best model.

This algorithm for model-based clustering allows mixed data including continuous, integer and categorical variables. Observations are assumed to be independent and coming from a mixture of g components. Each component is defined by their probability distribution function (pdf) such that:

$$f(x_i|g,\theta) = \sum_{k=1}^{g} \tau_k f_k(x_i|\alpha_k) \quad with \quad f_k(x_i|\alpha_k) = \prod_{j=1}^{d} f_{kj}(x_{ij}|\alpha_{kj}) \tag{4.2}$$

Taking in account n observations $x = (x_1....x_n)$. θ groups the model parameters, τ_k is the proportion represented by each component k such that it adds to 1 for the sum of τ_k for every component. f_k is the pdf of component k defined by all its parameters α_k. f_{kj} is then the pdf of variable j for component k defined by the parameter α_{kj}. The definition of space varies depending on the nature of the variable, so for continuous, integer and categorical variables the pdf(f_{kj}) and parameters (α_{kj}) will correspond to those of a Gaussian, Poisson or Multinomial distribution respectively (Marbac M et al., 2020).

Each variable is defined as relevant (with $\omega = 1$) or irrelevant (with $\omega = 0$) by a binary vector $\omega = (\omega_1....\omega_n)$. Then for a model defined by the parameter space $(m = g, \omega)$ we update the definition including this component to the probability distribution function for x_i such that:

$$f(x_i|m,\theta) = \prod_{j\in\Omega^c} f_{\omega j}(x_{ij}|\alpha_{\omega j}) \sum_{k=1}^{g} \tau_k \prod_{j\in\Omega} f_{kj}(x_{ij}|\alpha_{kj}) \tag{4.3}$$

where $\Omega = j : \omega_j = 1 and \Omega^c = 1...n \notin \Omega$.

To fit the model, a modified version of expectation maximization (EM) algorithm is used. *VarseLCM* uses Bayesian Information Criterion (BIC) and maximum likelihood inference simultaneously, the authors call this the penalized complete-data log-likelihood function:

$$l_{pen}(\Theta|m,x,z) = l(\theta|m,x,z) - (g-1)\frac{1}{2}lnn - \frac{1}{2}lnn\sum_{j=1}^{d} v_j(g\omega_j + 1 - \omega_j) \tag{4.4}$$

were vj is the number of parameters for the marginal distribution of variable j.

A random initialization for a fixed number of components is performed as first step and then the algorithm iterates between EM steps:

- Expectation step (E), were log likelihood is calculated for each observation given current model parameters.

$$\tau_{ik}^{[r]} := \frac{\tau_k^{[r-1]} \prod_{j=1}^{d} f_{kj}(x_{ij}|\alpha_{kj}^{[r-1]})}{\sum_{l=1}^{g} \tau_l^{[r-1]} \prod_{j=1}^{d} f_{lj}(x_{ij}|\alpha_{lj}^{r-1})} \tag{4.5}$$

- And then the maximization step (M) were model parameters and relevance of each variable are recalculated using posterior probabilities for each observations and maximum penalized log-likelihood estimators.

$$\omega_j^{[r]} = \begin{cases} 1, & if\ \triangle_j^{[r]} > 0 \\ 0, & otherwise \end{cases}; \quad \tau_k^{[r]} = \frac{n_k^{[r]}}{n} \quad \text{and} \quad \alpha_{jk}^{[r]} = \begin{cases} \alpha_{kj}^{\star[r]}, & if\ \omega_j^{[r]} = 1 \\ \widetilde{\alpha}_{kj}, & otherwise \end{cases} \tag{4.6}$$

were $[r]$ refers to each iteration and $\triangle_j$ is the difference between maximum penalized log-likelihood estimators when variable j is consider relevant ($\omega = 1$) vs. irrelevant ($\omega = 0$).

The algorithm stops when it converges to a local optima. Maximization of penalized log-likelihood is obtained after performing several random initializations for a fixed number of components for every number between g1 and gmax. BIC penalty is directly included within EM

algorithm easing model selection processes for large datasets in expense of assuming independence of variables.

Within the function other options criteria may be selected for mother selection. Given the number of observations and variables used for the cluster model we chose to use BIC criterion as found most adequate by the authors (Marbac M et al., 2020).

To evaluate cluster results, discriminative power and fitted distribution of most discriminative variables was evaluated by plots directly implemented on *VarselLCM* (Marbac M et al., 2019). Summary of the probabilities of misclassification for each cluster was also evaluated by plots as provided within the function results. Adjusted Rand Index (ARI) was used to evaluate agreement between partitions generated from both models using ARI function also available within *VarselLCM*.

To visualize the resulting partition, clusters were represented using a 3D scatter plot generated with *plotly* library (Sievert C, 2020) using the first 3 components of PCA applied to the numeric variables initially used as input to fit the clustering model. PCA analysis was performed using *prcomp* function from R-base stats functions.

Summary statistics, including model input variables and clinical and scanner data of interest, were described for every cluster. Cross tables were generated to describe joint and marginal relative frequencies for histology classes and corresponding clusters, and for overall stage and corresponding clusters. Correspondence analysis was performed using *ca* library and asymmetric representations were generated using principal coordinates to represent clusters and standard coordinates to represent histology or overall stage classes (Nenadic et al, 2007) . The clusters with highest inertia as evaluated by correspondence analysis, were further evaluated for difference in distribution of variables between the selected clusters, including both variables within and outside the model.

To evaluate association between survival and different clusters, Kaplan Meyer curves were estimated using Survival.time and deadstatus.event variables, and survival curves for each cluster were compared using log-rank test, again, using *Survival* and *survminer* packages (Therneau T, 2022; Kassambara A et al., 2021). The clusters with greatest and lowest median survival were further compared by describing and testing differences in distribution of variables both within and outside the model.

Level of significance was set to α 0.05. Adjusted α value according to Bonferroni correction was provided on every table where multiple tests are performed at the same time.

Chapter 5

Results

5.1 Data frame Dimensions and Data types

Along with PatientID 9 Clinical variables were available in the associated csv file including: Age, Gender, Clinical T Stage, Clinical N Stage, Clinical M Stage, Overall.Stage, Histology, Survival time and Dead status event. CT studies available were dated from 2004-09-27 to 2014-01-01. Patient 128 did not have a GTV-1 segmentation available so was excluded from the beginning as no feature extraction was possible.

After combining clinical data, selected DICOM metadata, and pyradiomics output a data frame with 421 observations and 1257 variables was obtained including:

- 9 Clinical variables
- 2 Dicom metadata variables
- 1246 radiomic features

All 1246 radiomic features were continuous variables.

5.2 Initial Exploratory Data Analysis

On an initial exploratory analysis we found 66 missing values. Histology values were missing in 42 subjects, age value was missing in 22 subjects, Overall.Stage in 1 subject and clinical T stage in 1 subject (Figure 5.1). No missing values were found for the rest of the clinical variables, selected DICOM metadata or radiomics features. Overall completeness rate was superior to 0.9 for every variable. Forty-two subjects with missing value for histology variable were excluded, leaving a total of 379 observations to continue with the analysis.

Var.name	Var.Type
age	num
clinical.T.Stage	Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<"5"
Clinical.N.Stage	Ord.factor w/ 5 levels "0"<"1"<"2"<"3"<"4"
Clinical.M.Stage	Ord.factor w/ 3 levels "0"<"1"<"3"
Overall.Stage	Ord.factor w/ 4 levels "I"<"II"<"IIIa"<"IIIb"
Histology	Factor w/ 4 levels "adenocarcinoma","large cell","squamous cell carcinoma","nos"
gender	Factor w/ 2 levels "female","male"
Survival.time	int
deadstatus.event	Factor w/ 2 levels "0","1"
Study.Date	Date, format: "2008-09-18"
Manufacturer	Factor w/ 3 levels "CMS Inc.","Plastimatch","SIEMENS"

Table 5.1: Table describing clinical and dicom metadata variables.

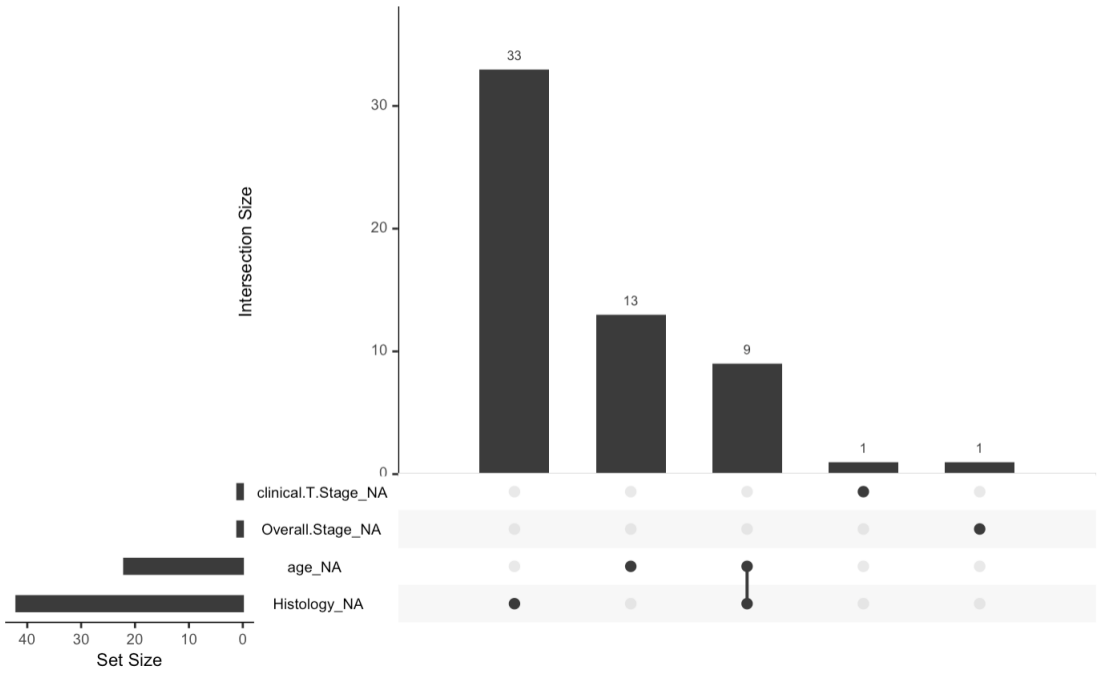


Figure 5.1: Barplot showing variables with missing values. Univariate missing value counts are displayed in the bottom left horizontal plot. Intersection size is represented with the main vertical barplot.

Regarding exploratory analysis for categorical variables we found a female/male ratio of 0.46 (119/260). The three major NSCLC histotypes were represented within the dataset; squamous cell carcinoma class predominated with152 observations (40%), followed by large

cell carcinoma (30%), NOS class (16%) and adenocarcinoma (13%). Regarding Overall.Stage variable distribution, we found stage IV was not represented in this dataset which forced us to center analysis on this group of non-metastatic disease patients; the highest proportion of patients fell within Overall stage IIIa and b stages. Some incoherent data entries were also found regarding clinical T, N and M categories. As previously mentioned, possible values for clinical.T.Stage category in lung cancer ranges from T1 to T4, and few cases were identified with value entry of T = 5. Same happened with Clinical.N.Stage = 4 and and some cases with Clinical M Stage = 3. Regarding scanner manufacturer model, 292 studies were obtained with a SIEMENS Biograph 40 scanner, and 86 studies with a CMS Inc. XiO. 1 CT study had Plastimach as Manufacturer value, which seemed to be an error as well as this is a software for image computation (Sharp et al., 2010). These findings are further detailed on Table 5.2 and Figure 5.2.

Histology	n	rate
adenocarcinoma	51	0.1345646
large cell	114	0.3007916
nos	62	0.1635884
squamous cell carcinoma	152	0.4010554

gender	n	rate
female	119	0.3139842
male	260	0.6860158

deadstatus.event	n	rate
0	43	0.1134565
1	336	0.8865435

Manufacturer	n	rate
CMS Inc.	86	0.2269129
Plastimatch	1	0.0026385
SIEMENS	292	0.7704485

Overall.Stage	n	rate
I	66	0.1741425
II	38	0.1002639
IIIa	108	0.2849604
IIIb	166	0.4379947
NA	1	0.0026385

clinical.T.Stage	n	rate
1	69	0.1820580
2	147	0.3878628
3	50	0.1319261
4	111	0.2928760
5	1	0.0026385
NA	1	0.0026385

Clinical.N.Stage	n	rate
0	140	0.3693931
1	21	0.0554090
2	134	0.3535620
3	81	0.2137203
4	3	0.0079156

Clinical.M.Stage	n	rate
0	375	0.9894459
3	4	0.0105541

Table 5.2: Absolute and relative frequency tables for each categorical variable.

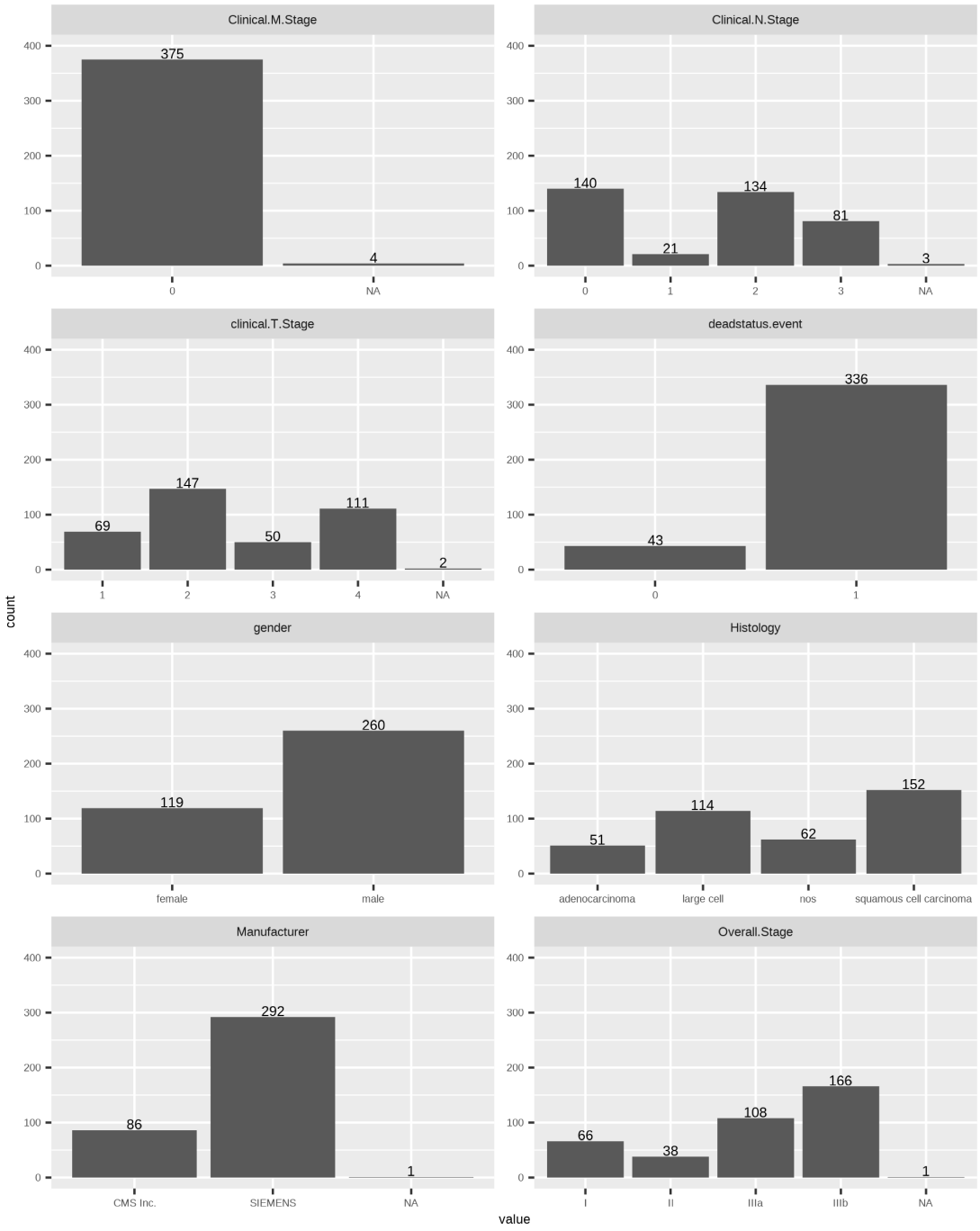


Figure 5.2: Barplots showing distribution of categorical variables.

All entries interpreted as erroneous were replaced with missing values for further imputation. As Clinical.M.Stage was mainly populated by M0 and missing values, this variable was then eliminated assuming cero variance.

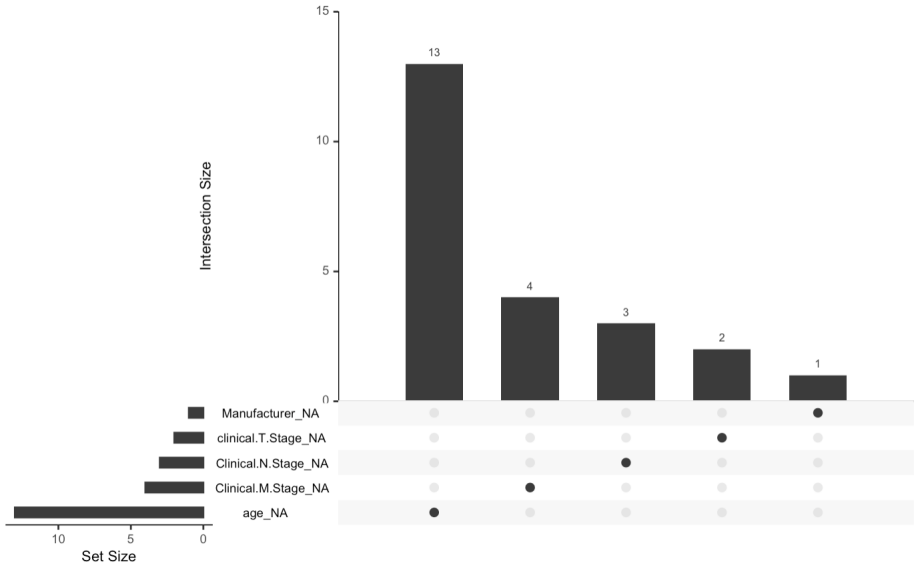


Figure 5.3: Missing value pattern after removing observations with histology missing values and imputing with missing values the incoherent clinical T/N/M and Manufacturer entries.

Regarding continuous clinical variables distribution, as expected, we can see age shows a relatively symmetric distribution while survival time shows a skewed distribution (Figure 5.4) Though there is an outlier observation for the youngest patient, it still seems a reasonable age, does not seem an error. Similar information is inferred from skewness and kurtosis values for these variables, near zero for age and greater than 1 for survival time (Table 5.3).

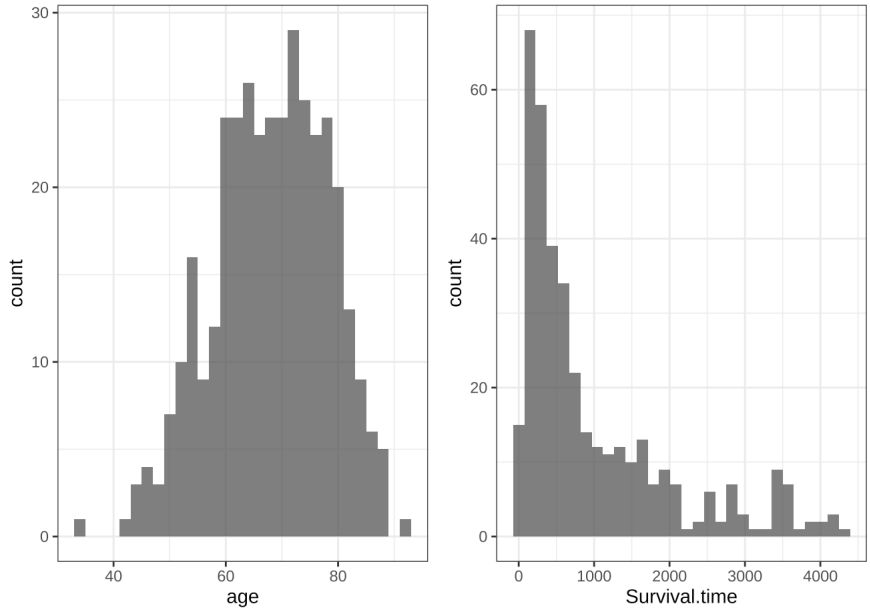


Figure 5.4: Histograms showing distribution of clinical continuous variables.

described_variables	n	na	mean	sd	se_mean	IQR	skewness	kurtosis
age	366	13	68.1016	10.14966	0.5305312	14.74875	-0.3099521	-0.3094659
Survival.time	379	0	983.6464	1020.99688	52.4450869	1126.00000	1.4764935	1.3393722

Table 5.3: Table describing summary statistics for age and survival time.

When analyzing continuous variable distribution with the Shapiro Wilk test, only 167 out of the 1248 variables showed results favoring underlying normal distribution taking in account alpha of 0.05 with Bonferroni correction. Given the high dimensionality of radiomic variables further analysis and univariate data visualization for these variables was performed after feature selection.

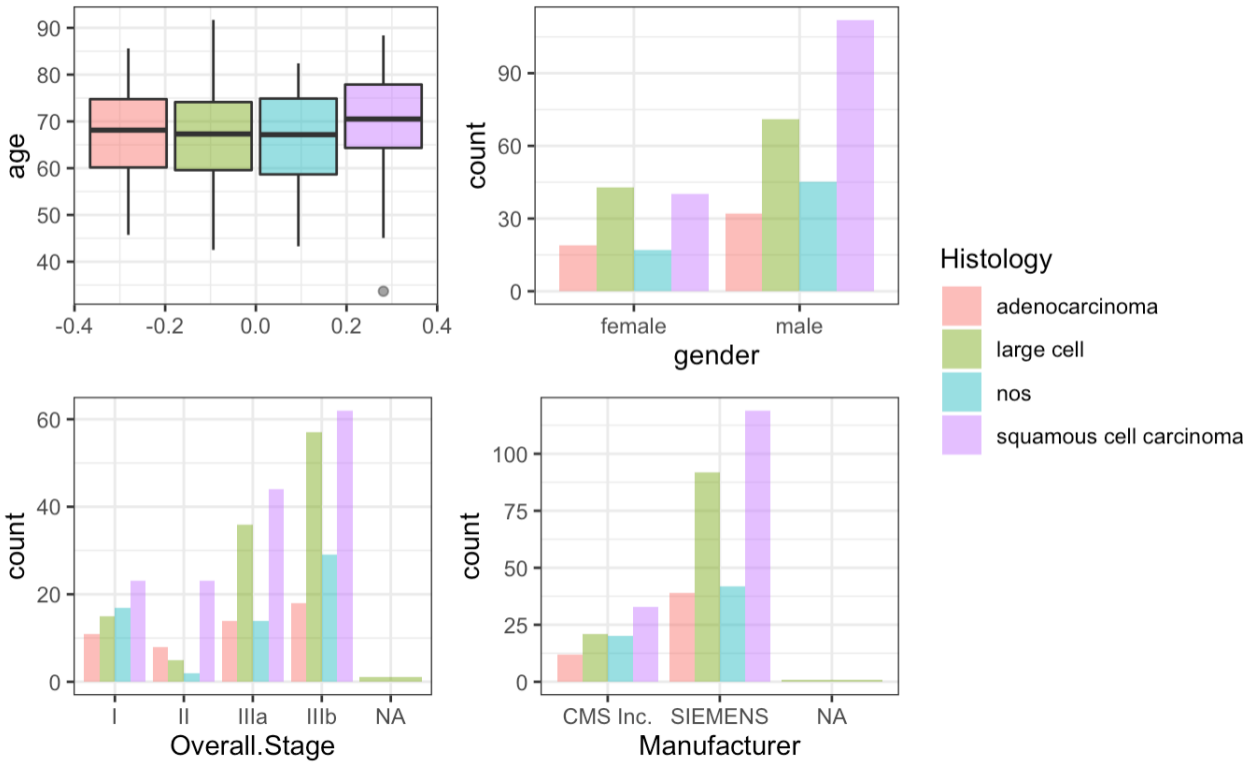


Figure 5.5: Boxplot and histograms showing the relationship between different variables and histology class.

Concerning relationship between histology and clinical/scanner variables of interest, taking in account corrected α, we found a significant relationship between histology and age (p value = 0.0076); patients with Squamous cell carcinoma showed a slightly older mean age than other histology groups. Regarding histology and the overall stage independence test showed p value = 0.0117 against the null hypothesis, though not outside the limits defined when we compare to

Bonferroni corrected α, we still took this potential relationship into account when comparing the results. We found no statistically significant association between histology and other categorical variables. Differences between groups can be further appreciated on Figure 5.5 and Table 5.4. As a relevant point we found independence between histology and Manufacturer classes and regarding the period of time during which studies were performed, we found all classes were represented along the years in favor of avoiding specific technique batch biases when evaluating this variable as target (Figure 5.6).

label	variable	Histology				test
		adenocarcinoma	large cell	nos	squamous cell carcinoma	
age	Min / Max	45.7 / 85.6	42.5 / 91.7	43.3 / 82.4	33.7 / 88.4	p value: 0.0076 (One-way analysis of means)
	Med [IQR]	68.1 [60.2;74.8]	67.3 [59.6;74.1]	67.2 [58.7;74.9]	70.5 [64.3;77.9]	
	Mean (std)	67.3 (9.6)	66.9 (10.2)	65.6 (10.3)	70.2 (9.9)	
	N (NA)	49 (2)	110 (4)	58 (4)	149 (3)	
gender	female	19 (15.97%)	43 (36.13%)	17 (14.29%)	40 (33.61%)	p value: 0.1574 (Pearson's Chi-squared test)
	male	32 (12.31%)	71 (27.31%)	45 (17.31%)	112 (43.08%)	
clinical.T.Stage	1	12 (17.39%)	18 (26.09%)	14 (20.29%)	25 (36.23%)	p value: 0.4042 (Pearson's Chi-squared test)
	2	22 (14.97%)	47 (31.97%)	23 (15.65%)	55 (37.41%)	
	3	9 (18.00%)	11 (22.00%)	7 (14.00%)	23 (46.00%)	
	4	8 (7.21%)	37 (33.33%)	17 (15.32%)	49 (44.14%)	
Clinical.N.Stage	0	18 (12.86%)	32 (22.86%)	27 (19.29%)	63 (45.00%)	p value: 0.1737 (Pearson's Chi-squared test)
	1_2	19 (12.26%)	53 (34.19%)	21 (13.55%)	62 (40.00%)	
	3	14 (17.28%)	28 (34.57%)	14 (17.28%)	25 (30.86%)	
Overall.Stage	I	11 (16.67%)	15 (22.73%)	17 (25.76%)	23 (34.85%)	p value: 0.0117 (Pearson's Chi-squared test)
	II	8 (21.05%)	5 (13.16%)	2 (5.26%)	23 (60.53%)	
	IIIa	14 (12.96%)	36 (33.33%)	14 (12.96%)	44 (40.74%)	
	IIIb	18 (10.84%)	57 (34.34%)	29 (17.47%)	62 (37.35%)	
Manufacturer	CMS Inc.	12 (13.95%)	21 (24.42%)	20 (23.26%)	33 (38.37%)	p value: 0.2199 (Pearson's Chi-squared test)
	SIEMENS	39 (13.36%)	92 (31.51%)	42 (14.38%)	119 (40.75%)	

Table 5.4: Table showing the relationship between different variables and histology class, and the result of different tests to evaluate independence. Table-wise Bonferroni corrected $\alpha = 0.008$.

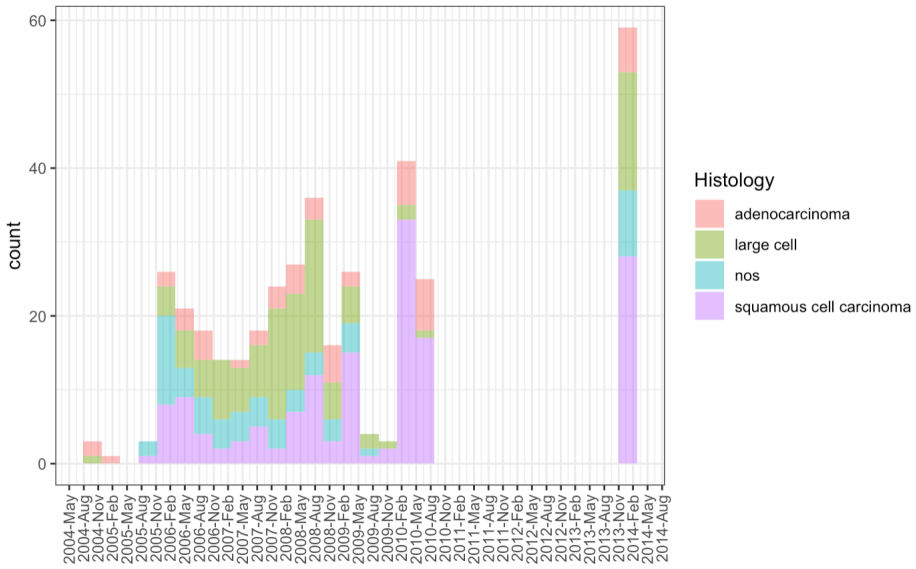


Figure 5.6: Histogram representing the distribution of different histology classes along different months and years during which imaging data was performed.

Regarding Overall stage we found a significant relationship with age ($p < 0.0001$), with patients with overall stage I showing the oldest mean age and patients with overall stage IIIb showing the youngest mean age. No significant association was found between overall stage and gender, or overall stage and scanner manufacturer model (Table 5.5). As known from theory, overall stage is dependent on clinical T/N/M variables, we found a Spearman correlation coefficient of 0.58 for overall stage and clinical T stage, and of 0.51 for overall stage and clinical N stage. T and N categories did not show an important correlation (Spearman c.: -0.08).

label	variable	Overall.Stage				test
		I	II	IIIa	IIIb	
age	Min / Max	51.7 / 91.7	50.2 / 87.0	33.7 / 85.1	42.5 / 88.4	p value: <0.0001 (Kruskal-Wallis rank sum test)
	Med [IQR]	75.7 [68.5;80.2]	73.4 [70.2;78.6]	67.1 [60.3;74.1]	65.8 [59.9;72.8]	
	Mean (std)	74.1 (9.0)	72.9 (9.1)	66.7 (10.4)	65.8 (9.4)	
	N (NA)	62 (4)	35 (3)	106 (2)	162 (4)	
gender	female	19 (15.97%)	9 (7.56%)	38 (31.93%)	53 (44.54%)	p value: 0.5734 (Pearson's Chi-squared test)
	male	47 (18.15%)	29 (11.20%)	70 (27.03%)	113 (43.63%)	
Manufacturer	CMS Inc.	22 (25.58%)	6 (6.98%)	21 (24.42%)	37 (43.02%)	p value: 0.0987 (Pearson's Chi-squared test)
	SIEMENS	43 (14.78%)	32 (11.00%)	87 (29.90%)	129 (44.33%)	

Table 5.5: Table showing the relationship between different variables and overall stage, and the result of different tests to evaluate independency. Table-wise Bonferroni corrected $\alpha = 0.0125$.

When evaluating survival for the whole group with Kaplan-Meyer curves, we found a median

survival of 573 days (95%CI: 492-660 days). When evaluating the relationship between survival and clinical and scanner variables of interest, we found no significant differences in survival curves between histology groups (p = 0.33), gender (p = 0.11) or even Overall stage groups (p = 0.52). On the other hand, we did find a difference between survival curves of observations corresponding to scanners from different manufacturers. Though 95% CI between median survival overlaps (CMS Inc. group 444 days CI:325-597 vs. SIEMENS 617 days CI:522-704), we found a p value of 0.002 when testing for difference between both curves with log-rank test. This could add a major bias to interpreting model results regarding survival analysis as target variable. When performing radiomic features selection we should exclude any variables related to scanner manufacturer with the goal of reducing bias from this source.

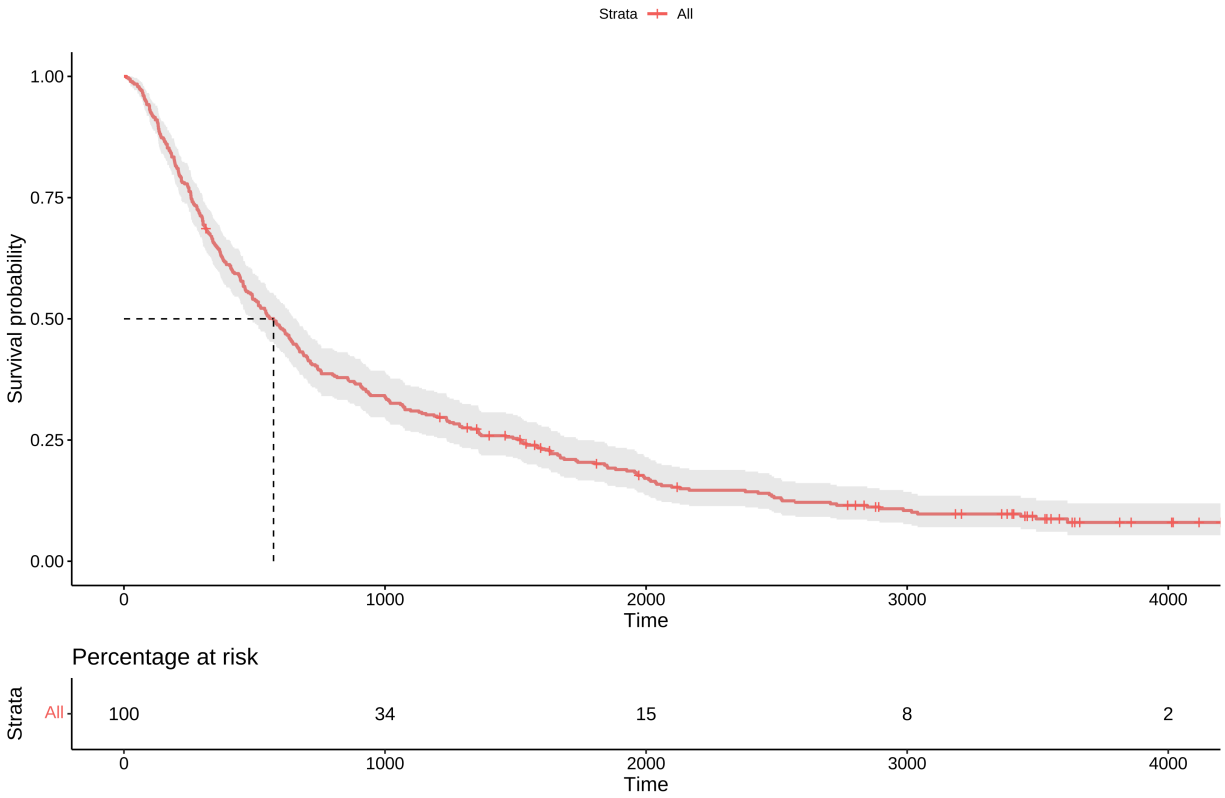


Figure 5.7: Kaplan Meyer plot showing survival probability over time in days for the whole group of patients. Median survival is indicated with the dotted line.

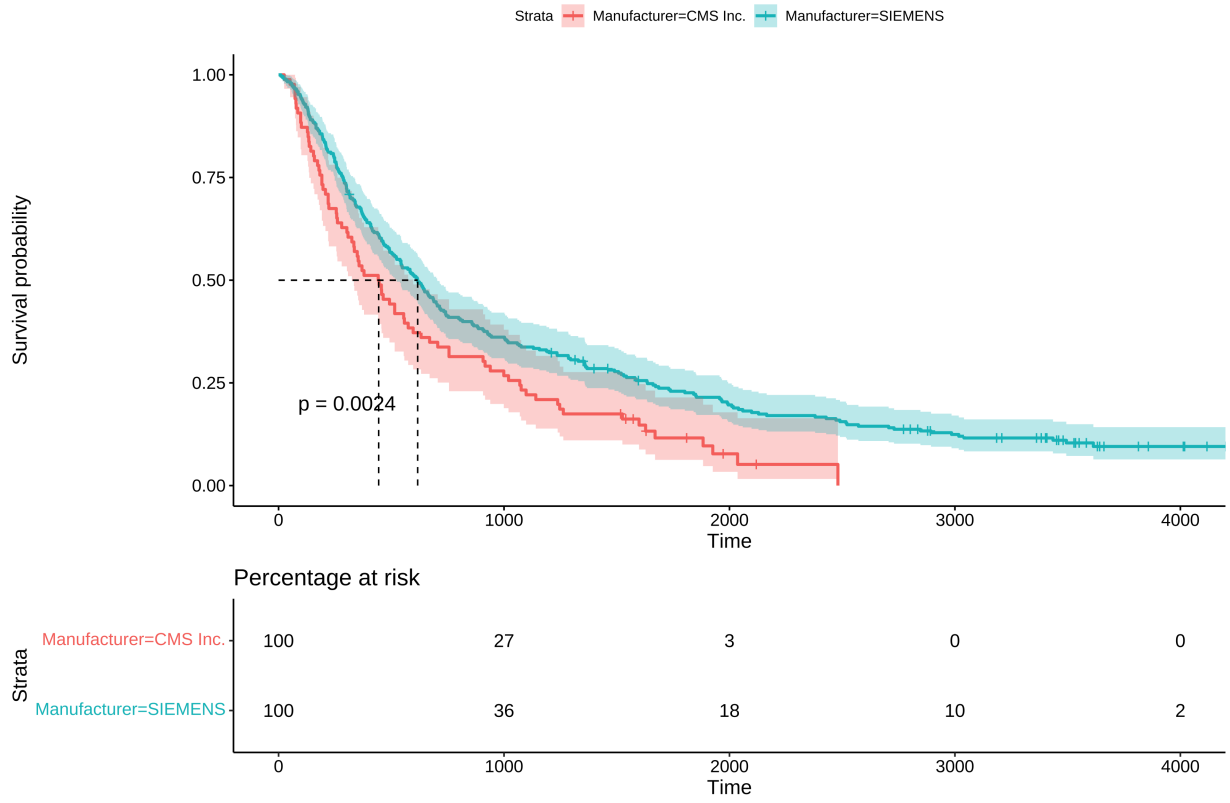


Figure 5.8: Kaplan Meyer plot showing survival probability over time in days given different scanner Manufacturer model used for CT exam. Median survival for each group is indicated with the dotted line and p value corresponding to the log-rank test between curves is indicated on the plot. On the bottom we can see the percentage of patients at risk at each selected time point for each group.

5.2.1 Missing values and outliers imputation

After imputation of missing values with *missForest* algorithm, we obtained an Out-of-bag error of 4.419509e-12 for normalized mean squared error, and 2.141340e-01 proportion of falsely classified for continuous and categorical variables respectively.

Two observations were identified as outliers ("LUNG1-027" "LUNG1-069") after applying HDoutliers multivariate outlier detection (Figure 5.9).

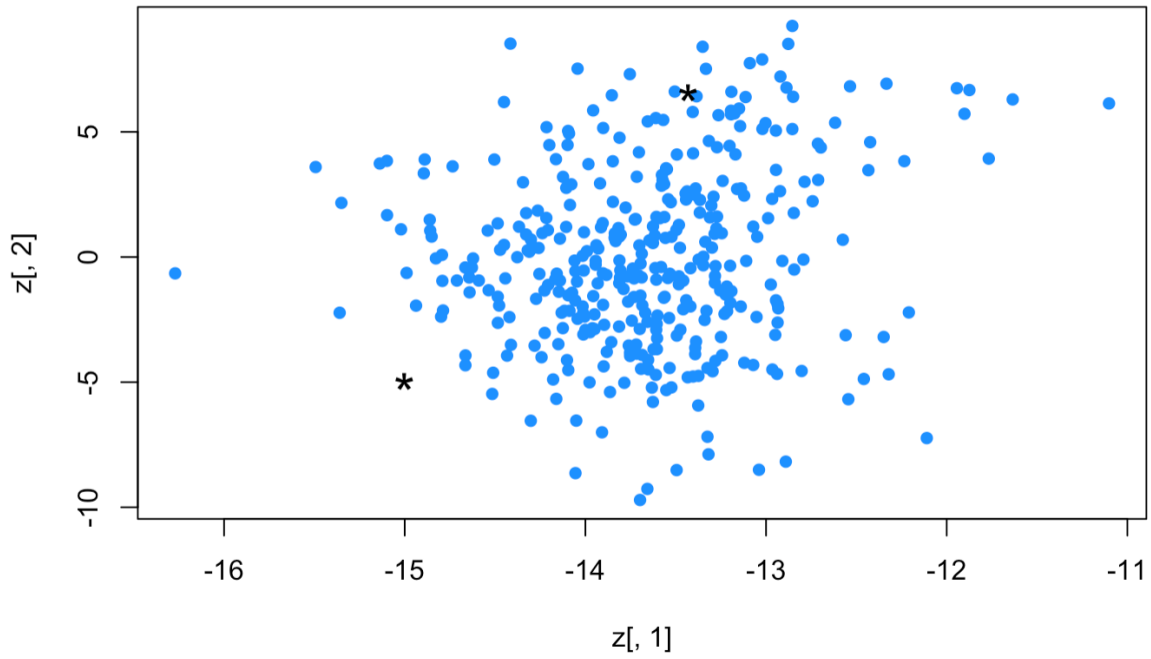


Figure 5.9: Scatter plot displaying the results obtained when performing multivariate outlier detection with HDoutliers method.

Both observations were excluded from the dataset, leaving 377 observations left for further analysis. On table 5.6 we can see the original number of observations and the different steps that lead to the final number of observations used for further analysis.

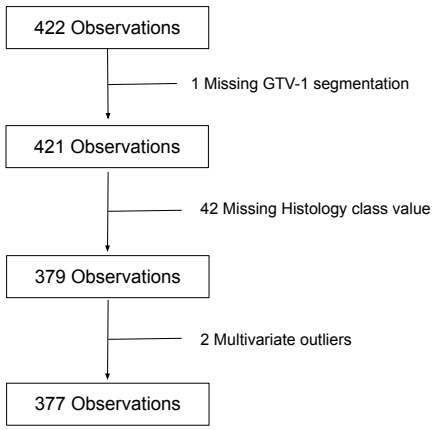


Table 5.6: Flowchart summarizing number of observations excluded and main reasons for it.

5.2.2 Radiomic features selection, transformation and further exploratory analysis

As for radiomic variables, we can see the high dimensionality (1246 vars) and high correlation of radiomic variables, as evaluated by pearson correlation coefficient, represented with a heatmap plot in figure 5.10.

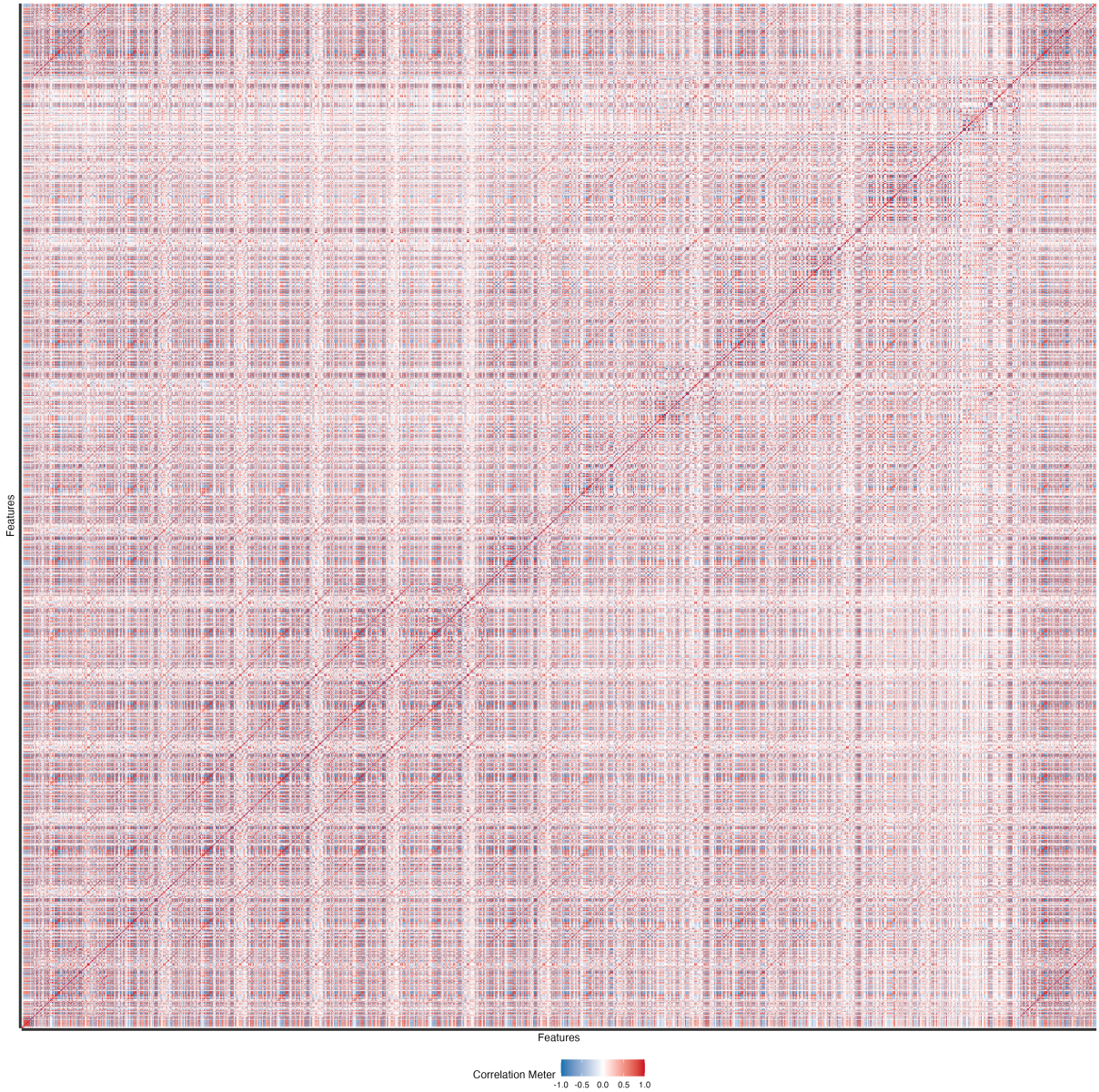


Figure 5.10: Heatmap displaying correlation between 1246 continuous radiomic variables by pearson correlation coefficient.

After verifying the relationship of different radiomic features to the scanner manufacturer

model, 714 features were removed with 532 radiomic features remaining on the dataset. No near cero radiomic features were detected for this subset. After joining results from MRMR feature selection as evaluated for both Histology and Survival variables, a total of 36 radiomic features were selected. Four features were selected in both cases including:

- log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis,
- wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis,
- wavelet.HHH_glcm_ClusterProminence,
- log.sigma.4.0.mm.3D_glszm_ZoneVariance.

Five features were additionally filtered out for redundancy, as previously defined in methodology section (Pearson correlation coefficient above 0.95) including:

- wavelet.HHL_glrlm_ShortRunHighGrayLevelEmphasis,
- log.sigma.5.0.mm.3D_glszm_LargeAreaEmphasis,
- log.sigma.4.0.mm.3D_glszm_LargeAreaEmphasis,
- log.sigma.4.0.mm.3D_glszm_ZoneVariance,
- log.sigma.4.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis.

A total of 31 radiomic features were retained for further analysis. On figure 5.11 we can see a heatmap representing pearson correlation coefficient values for the selected radiomic features and in table 5.7 we can see summary statistics for selected variables.

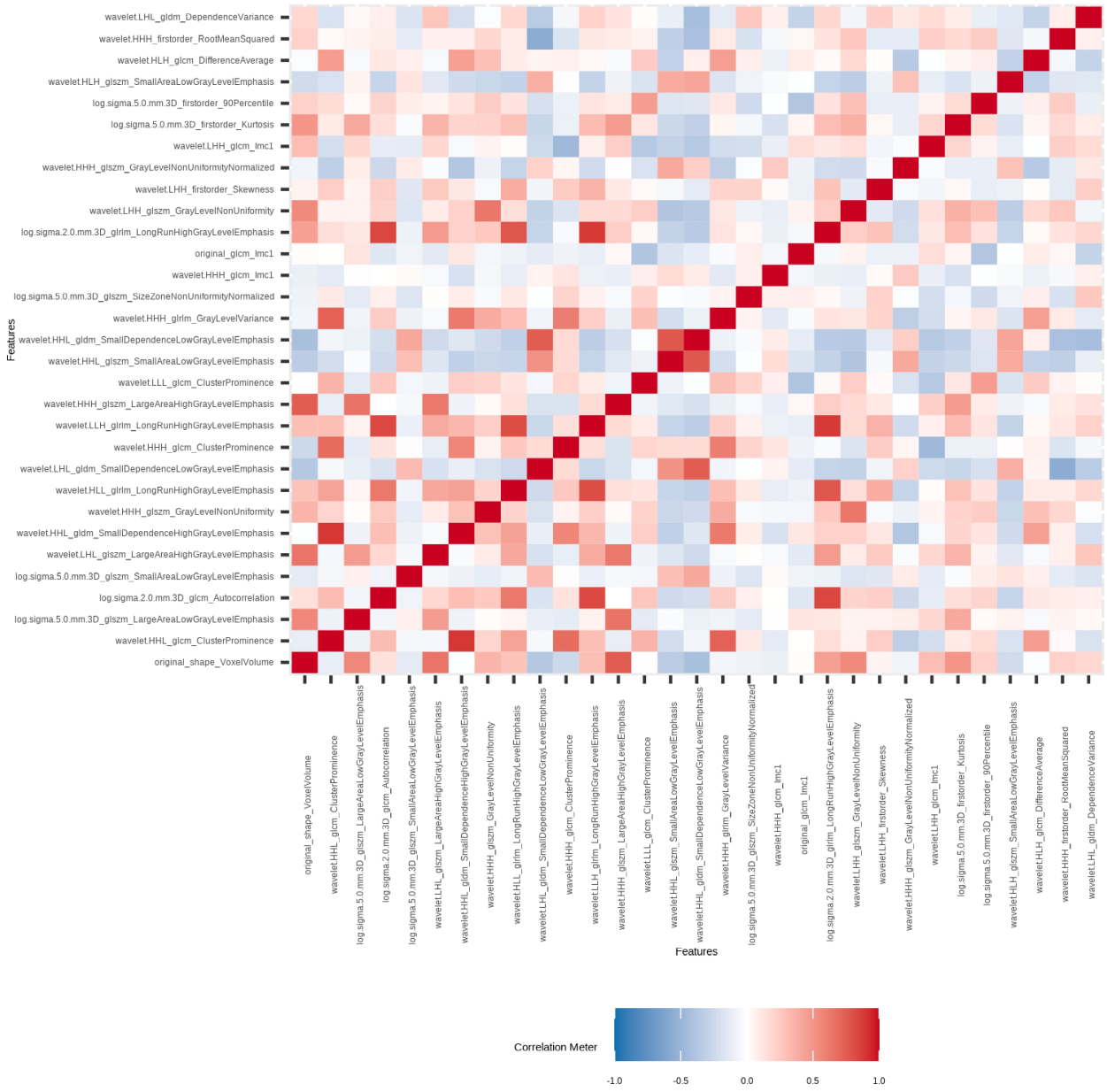


Figure 5.11: Heatmap representing correlation coefficients between the 31 selected radiomic features.

described_variables	mean	sd	IQR	skewness	kurtosis
original_shape_VoxelVolume	7.514392e+04	9.164034e+04	9.800700e+04	2.0851762	4.8456596
wavelet.HHL_glcm_ClusterProminence	7.665988e+01	1.454862e+02	7.228063e+01	8.3455080	103.1677055
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	1.019452e+04	4.332770e+04	3.566937e+03	9.0663774	102.3326825
log.sigma.2.0.mm.3D_glcm_Autocorrelation	3.906822e+02	4.635002e+02	1.964977e+02	7.0022832	64.9811527
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	2.035400e-03	2.590600e-03	8.596000e-04	6.2617573	50.2486161
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	1.386738e+08	4.014961e+08	9.273493e+07	6.7919498	63.6744640
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	4.974261e+00	7.463963e+00	5.531913e+00	10.3666533	156.5331460
wavelet.HHH_glszm_GrayLevelNonUniformity	6.279034e+00	1.078154e+01	4.444444e+00	5.5080663	37.7984974
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	2.202244e+03	3.333935e+03	1.541908e+03	6.3181067	54.5274282
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	6.336000e-04	9.168000e-04	4.890000e-04	4.8892521	35.0606807
wavelet.HHH_glcm_ClusterProminence	5.201669e-01	2.434530e-02	2.194800e-02	7.6931019	100.7114041
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	8.932846e+02	1.224833e+03	6.660940e+02	6.0260294	44.3592347
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	2.029759e+09	5.437150e+09	1.779997e+09	6.6191390	54.5731854
wavelet.LLL_glcm_ClusterProminence	4.011467e+07	5.602678e+07	3.576071e+07	4.2809018	26.1167309
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	1.579710e-02	1.467100e-02	1.358980e-02	3.4688959	22.4729366
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	1.426500e-03	1.429000e-03	1.121800e-03	2.3362021	6.1842749
wavelet.HHH_glrlm_GrayLevelVariance	2.508655e-01	2.168600e-03	7.639000e-04	4.8123946	29.9377695
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	1.295846e-01	2.895060e-02	3.887950e-02	0.0035690	0.1889044
wavelet.HHH_glcm_Imc1	-1.806540e-02	3.470200e-03	4.504700e-03	-1.0817541	1.2105680
original_glcm_Imc1	-2.123565e-01	4.045460e-02	5.235530e-02	-0.4524797	1.5201208
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	2.179683e+03	3.172775e+03	2.061257e+03	5.0438332	34.4480218
wavelet.LHH_glszm_GrayLevelNonUniformity	1.303759e+02	1.629441e+02	1.371509e+02	3.2948811	16.5711634
wavelet.LHH_firstorder_Skewness	9.730470e-02	2.125665e-01	1.816981e-01	1.0791018	8.8925641
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	4.580535e-01	1.130034e-01	1.953485e-01	-0.0644575	-0.7458533
wavelet.LHH_glcm_Imc1	-5.209010e-02	9.000900e-03	1.068450e-02	-1.0971085	2.1923760
log.sigma.5.0.mm.3D_firstorder_Kurtosis	3.690143e+00	1.856490e+00	1.379258e+00	3.6248597	21.7097063
log.sigma.5.0.mm.3D_firstorder_90Percentile	1.147649e+01	7.613904e+01	6.642061e+01	0.5149007	3.4489701
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	7.941480e-02	6.270540e-02	4.685890e-02	2.1255940	4.6026079
wavelet.HLH_glcm_DifferenceAverage	4.690893e-01	4.681940e-02	5.090000e-02	1.7774220	5.4619428
wavelet.HHH_firstorder_RootMeanSquared	9.999963e+02	2.351280e-02	1.363290e-02	-3.6488720	19.5431225
wavelet.LHL_gldm_DependenceVariance	1.432114e+01	7.129074e+00	1.065779e+01	0.5921853	-0.2546894

Table 5.7: Univariate summary statistics for selected radiomic variables.

As we can appreciate on the table, most radiomic variables show high values for skewness and kurtosis, with some isolated exceptions as for those where previous normalization is already stated on the variable name. As we can see both in summary statistics and histograms, most variables are right-skewed and could benefit for log transformation. In figures 5.12 and 5.13 we can see qqplots for selected radiomic variables, before and after log transformation.

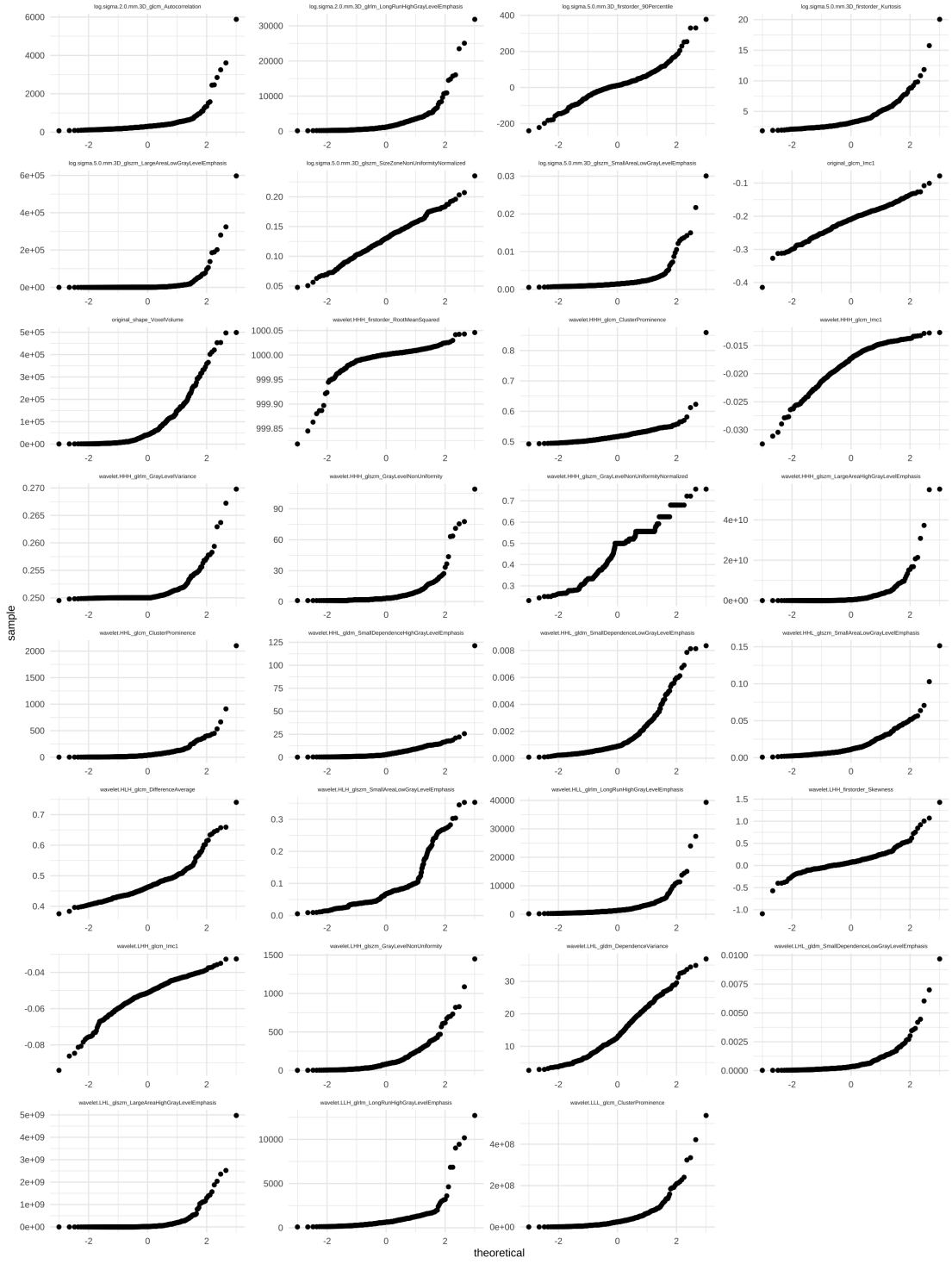


Figure 5.12: qqplots showing univariate radiomic observations against a normal distribution before log transformation.

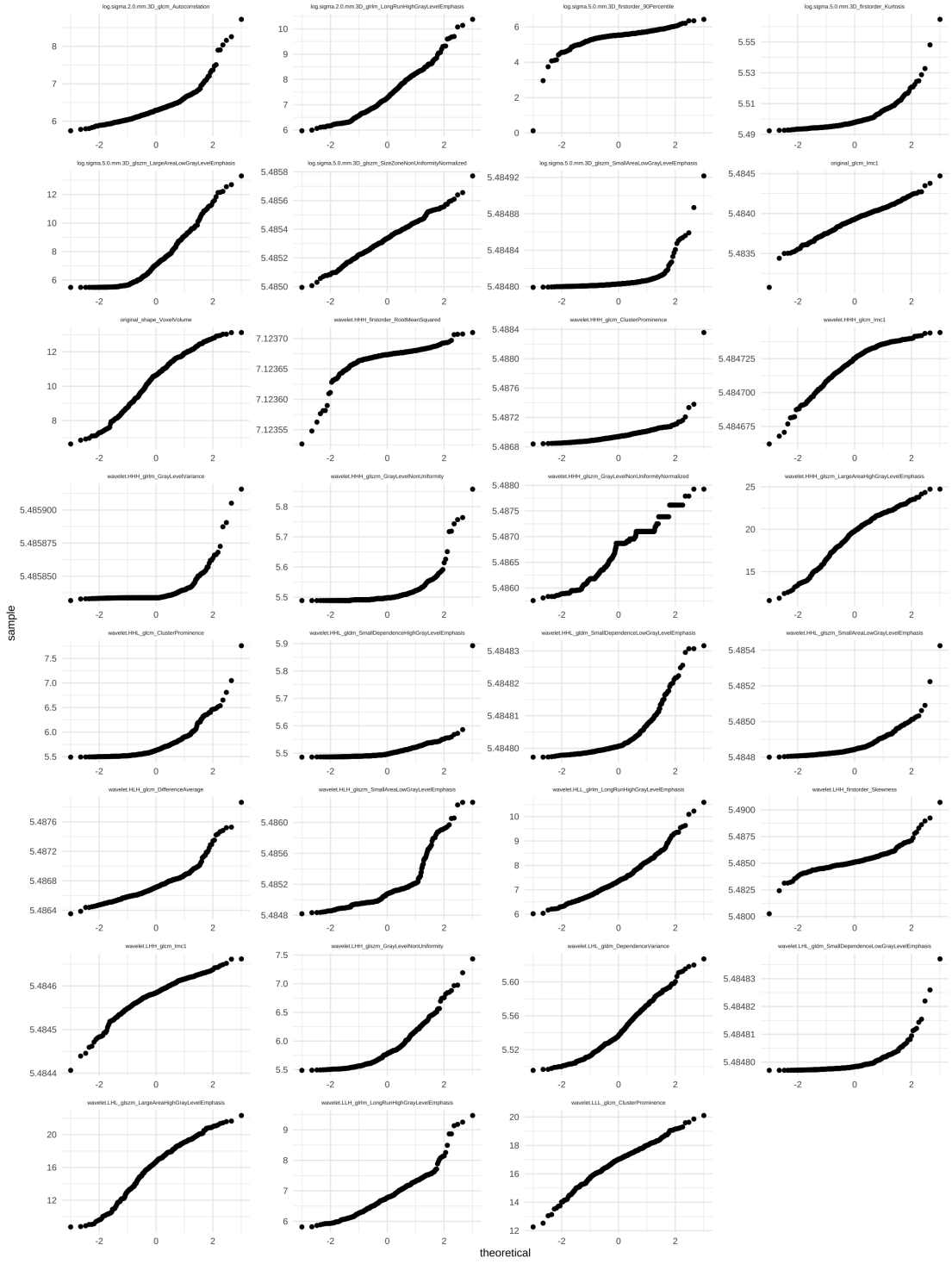


Figure 5.13: qqplots showing univariate radiomic observations against a normal distribution after log transformation.

When evaluating individual relationships of selected radiomic variables to histology class,

texture based features obtained after applying different wavelet filters were found as the most significantly related, though in the limit when we take into account table-wise Bonferroni corrected α. These findings are detailed in table 5.8. Distribution of most related radiomic features are represented with boxplots against histology class, on figure 5.14. Taking in account potential relationship between histology and Overall stage, relationship between selected radiomic variables and different Overall stage categories was also evaluated. Though texture features dominate as well, first order features as Kurtosis and shape features as voxel volume are also within the top list of related features for Overall.Stage categories (table 5.9, figure 5.15). Though most related radiomic features are different between both tables we can find some common features under 0.05 α level as wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis, wavelet.LHL_gldm _SmallDependenceLowGrayLevelEmphasis and wavelet.HHL_gldm_SmallDependenceLowGray LevelEmphasis.

label	variable	adenocarcinoma	large cell	nos	squamous cell carcinoma	pval
wavelet.LHL_gldm_DependenceVariance	Mean (std)	5.541 (0.028)	5.536 (0.024)	5.537 (0.024)	5.550 (0.030)	0.0012
wavelet.HHH_glcm_ClusterProminence	Mean (std)	5.487 (2e-04)	5.487 (6e-05)	5.487 (7e-05)	5.487 (7e-05)	0.0023
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	5.508 (0.057)	5.508 (0.020)	5.503 (0.017)	5.502 (0.017)	0.0032
wavelet.HHH_glcm_Imc1	Mean (std)	5.485 (1e-05)	5.485 (1e-05)	5.485 (2e-05)	5.485 (1e-05)	0.0145
wavelet.HHL_glcm_ClusterProminence	Mean (std)	5.700 (0.358)	5.748 (0.248)	5.692 (0.222)	5.703 (0.254)	0.0348
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	5.487 (2e-04)	5.487 (2e-04)	5.487 (2e-04)	5.487 (2e-04)	0.0384
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	5.505 (0.020)	5.514 (0.047)	5.501 (0.015)	5.511 (0.044)	0.0529
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	7.348 (0.914)	7.444 (0.670)	7.351 (0.642)	7.546 (0.728)	0.1077
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	5.485 (5e-06)	5.485 (5e-06)	5.485 (3e-06)	5.485 (3e-06)	0.1230
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	6.827 (0.697)	6.748 (0.523)	6.769 (0.545)	6.887 (0.565)	0.1665
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	5.487 (0.001)	5.487 (4e-04)	5.487 (4e-04)	5.487 (5e-04)	0.1696
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	5.451 (0.279)	5.510 (0.372)	5.500 (0.307)	5.428 (0.578)	0.1705
wavelet.LHH_firstorder_Skewness	Mean (std)	5.485 (0.001)	5.485 (0.001)	5.485 (0.001)	5.485 (0.001)	0.1760
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	5.485 (9e-05)	5.485 (4e-05)	5.485 (7e-05)	5.485 (5e-05)	0.2189
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	5.485 (6e-06)	5.485 (6e-06)	5.485 (7e-06)	5.485 (6e-06)	0.2246
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	5.485 (9e-06)	5.485 (1e-05)	5.485 (8e-06)	5.485 (1e-05)	0.2504
wavelet.LLL_glcm_ClusterProminence	Mean (std)	16.607 (1.106)	16.995 (1.184)	16.928 (1.170)	16.812 (1.337)	0.2752
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	5.486 (1e-05)	5.486 (9e-06)	5.486 (5e-06)	5.486 (1e-05)	0.2819
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	15.969 (3.046)	16.098 (2.874)	15.832 (3.055)	16.495 (2.887)	0.2850
wavelet.LHH_glcm_Imc1	Mean (std)	5.485 (3e-05)	5.485 (3e-05)	5.485 (5e-05)	5.485 (4e-05)	0.2863
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	5.820 (0.299)	5.894 (0.358)	5.825 (0.312)	5.848 (0.315)	0.4377
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	7.124 (2e-05)	7.124 (2e-05)	7.124 (2e-05)	7.124 (2e-05)	0.5500
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	5.500 (0.009)	5.500 (0.007)	5.500 (0.011)	5.499 (0.006)	0.5760
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	5.485 (3e-04)	5.485 (2e-04)	5.485 (3e-04)	5.485 (3e-04)	0.6286
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	6.310 (0.402)	6.348 (0.362)	6.344 (0.396)	6.362 (0.372)	0.6306
original_glcm_Imc1	Mean (std)	5.484 (2e-04)	5.484 (2e-04)	5.484 (2e-04)	5.484 (2e-04)	0.6505
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	7.361 (0.926)	7.373 (0.770)	7.352 (0.849)	7.439 (0.809)	0.7444
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	7.441 (1.803)	7.272 (1.560)	7.350 (1.858)	7.386 (1.569)	0.8484
original_shape_VoxelVolume	Mean (std)	10.365 (1.582)	10.450 (1.436)	10.270 (1.576)	10.472 (1.394)	0.8789
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	5.485 (1e-04)	5.485 (1e-04)	5.485 (1e-04)	5.485 (1e-04)	0.9716
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	19.215 (2.854)	19.309 (2.556)	19.237 (2.865)	19.349 (2.569)	0.9930

Table 5.8: Table showing the relationship between selected radiomic features and histology classes. Kruskal-Wallis rank sum test and One-way analysis of means was performed as appropriate to test difference in radiomic feature distribution between different histology classes, resulting p-value is included on the table. Table-wise Bonferroni corrected $\alpha = 0.0016$.

label	variable	I	II	IIIa	IIIb	pval
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	15.346 (2.887)	17.138 (2.686)	15.745 (2.779)	16.613 (2.991)	0.0002
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	7.214 (0.928)	7.616 (0.877)	7.235 (0.715)	7.519 (0.794)	0.0003
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	6.716 (0.668)	7.021 (0.620)	6.720 (0.514)	6.876 (0.536)	0.0004
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	5.485 (3e-06)	5.485 (3e-06)	5.485 (3e-06)	5.485 (5e-06)	0.0006
original_shape_VoxelVolume	Mean (std)	9.980 (1.427)	10.766 (1.461)	10.240 (1.404)	10.628 (1.462)	0.0011
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	18.651 (2.498)	19.903 (2.680)	18.926 (2.611)	19.666 (2.648)	0.0012
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	5.485 (5e-06)	5.485 (5e-06)	5.485 (6e-06)	5.485 (6e-06)	0.0022
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	5.499 (0.007)	5.500 (0.006)	5.500 (0.010)	5.500 (0.006)	0.0028
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	7.369 (0.822)	7.673 (0.884)	7.322 (0.645)	7.531 (0.685)	0.0048
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	6.903 (1.486)	7.571 (1.650)	7.286 (1.735)	7.525 (1.616)	0.0118
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	5.347 (0.423)	5.537 (0.287)	5.493 (0.351)	5.482 (0.534)	0.0199
wavelet.LLL_glcm_ClusterProminence	Mean (std)	17.136 (1.038)	16.772 (1.281)	16.992 (1.116)	16.681 (1.350)	0.0297
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	5.808 (0.317)	5.885 (0.291)	5.826 (0.347)	5.885 (0.322)	0.0418
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	5.505 (0.036)	5.513 (0.044)	5.511 (0.046)	5.510 (0.034)	0.1033
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	5.487 (0.001)	5.487 (5e-04)	5.487 (5e-04)	5.487 (5e-04)	0.1089
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	5.485 (3e-04)	5.485 (2e-04)	5.485 (3e-04)	5.485 (3e-04)	0.1283
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	5.485 (5e-05)	5.485 (1e-04)	5.485 (6e-05)	5.485 (5e-05)	0.1284
wavelet.LHL_gldm_DependenceVariance	Mean (std)	5.539 (0.026)	5.550 (0.030)	5.538 (0.025)	5.544 (0.029)	0.1337
wavelet.LHH_firstorder_Skewness	Mean (std)	5.485 (0.001)	5.485 (0.001)	5.485 (0.001)	5.485 (0.001)	0.2025
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	7.124 (2e-05)	7.124 (2e-05)	7.124 (1e-05)	7.124 (2e-05)	0.2119
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	6.342 (0.440)	6.411 (0.389)	6.307 (0.349)	6.363 (0.364)	0.2469
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	5.485 (1e-04)	5.485 (1e-04)	5.485 (1e-04)	5.485 (1e-04)	0.2482
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	5.485 (7e-06)	5.485 (1e-05)	5.485 (7e-06)	5.485 (1e-05)	0.2536
wavelet.HHL_glcm_ClusterProminence	Mean (std)	5.732 (0.229)	5.750 (0.402)	5.707 (0.253)	5.704 (0.246)	0.3005
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	5.486 (7e-06)	5.486 (1e-05)	5.486 (1e-05)	5.486 (6e-06)	0.3356
wavelet.LHH_glcm_Imc1	Mean (std)	5.485 (4e-05)	5.485 (4e-05)	5.485 (4e-05)	5.485 (4e-05)	0.3811
wavelet.HHH_glcm_ClusterProminence	Mean (std)	5.487 (7e-05)	5.487 (2e-04)	5.487 (9e-05)	5.487 (6e-05)	0.4075
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	5.487 (2e-04)	5.487 (2e-04)	5.487 (2e-04)	5.487 (2e-04)	0.4239
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	5.505 (0.017)	5.512 (0.065)	5.503 (0.016)	5.504 (0.020)	0.4650
wavelet.HHH_glcm_Imc1	Mean (std)	5.485 (2e-05)	5.485 (1e-05)	5.485 (1e-05)	5.485 (1e-05)	0.7165
original_glcm_Imc1	Mean (std)	5.484 (1e-04)	5.484 (2e-04)	5.484 (2e-04)	5.484 (2e-04)	0.8293

Table 5.9: Table showing the relationship between selected radiomic features and Overall stage categories. Kruskal-Wallis rank sum test and One-way analysis of means was performed as appropriate to test difference in radiomic feature distribution between different Overall stage categories, resulting p-value is included on the table. Table-wise Bonferroni corrected $\alpha = 0.0016$.

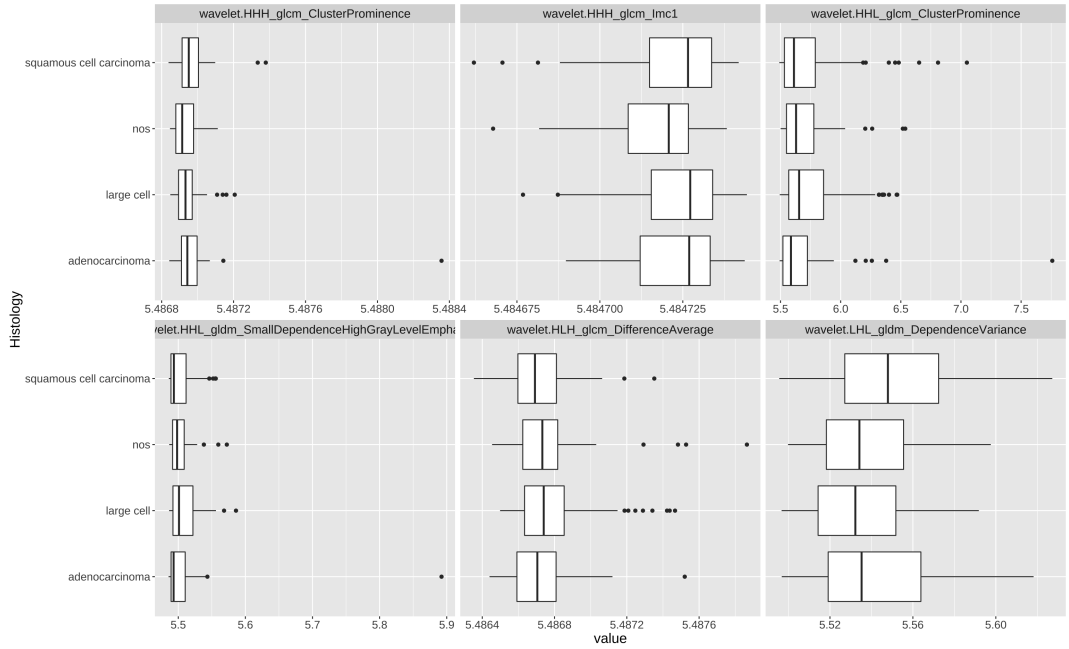


Figure 5.14: Most significantly associated features are displayed using boxplots against histology class.

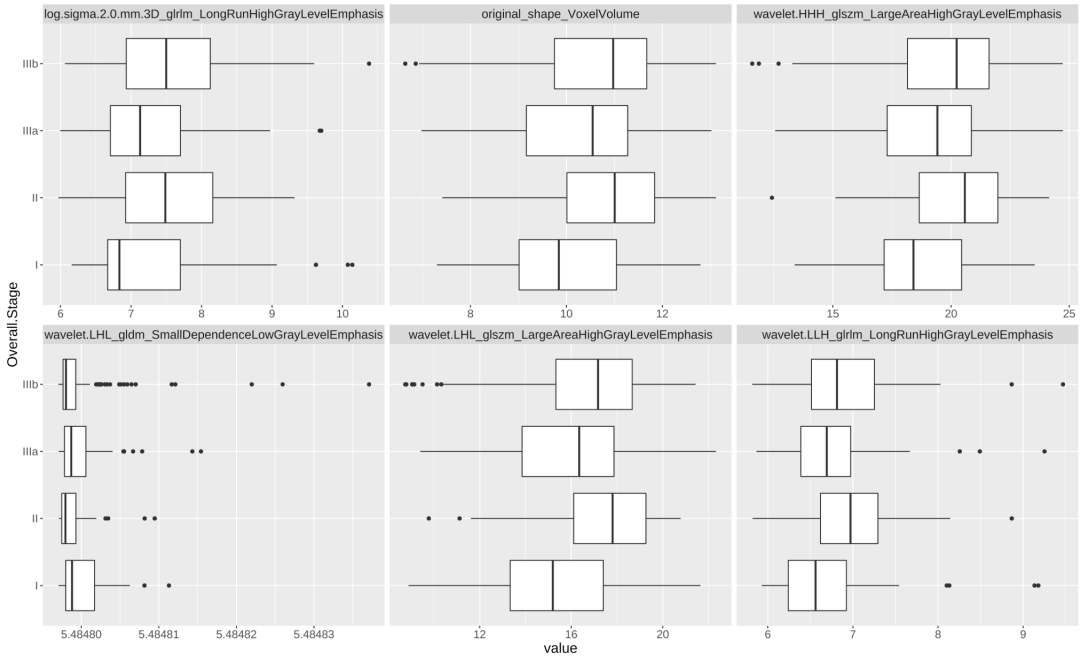


Figure 5.15: Most significantly associated features are displayed using boxplots against Overall stage class.

After fitting cox models to evaluate individual relationship of selected radiomic features to

survival, again many texture based features obtained after applying different log or wavelet filters found as significantly related and, as for Overall stage categories, first order features as Kurtosis and shape features as voxel volume were also included in the top list. Top list of related features shows many more coincidences between survival and overall stage analysis. These findings are detailed in table 5.10.

	beta	wald.test	p.value
wavelet.HHH_glszm_GrayLevelNonUniformity	5.2	19	0.000016
log.sigma.5.0.mm.3D_firstorder_Kurtosis	27	17	0.000043
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	0.29	17	0.000046
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	0.36	13	0.00024
wavelet.LHH_glszm_GrayLevelNonUniformity	0.62	13	0.00032
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	0.12	13	0.00035
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	0.27	11	0.00076
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	0.067	11	0.001
original_shape_VoxelVolume	0.13	10	0.0014
log.sigma.2.0.mm.3D_glcm_Autocorrelation	0.41	7.7	0.0054
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	0.055	6.2	0.013
wavelet.LHL_gldm_DependenceVariance	4.4	4.9	0.026
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	-440	4	0.044
wavelet.HHH_glcm_Imc1	-7000	3.4	0.064
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	-19000	3.2	0.073
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	-190	2.8	0.096
wavelet.HHH_glrlm_GrayLevelVariance	9100	2.6	0.11
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	-25000	2.1	0.14
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	-1600	2	0.16
wavelet.LHH_firstorder_Skewness	90	1.8	0.18
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	2	1.6	0.21
wavelet.LHH_glcm_Imc1	1800	1.1	0.29
wavelet.HHL_glcm_ClusterProminence	0.2	1	0.31
log.sigma.5.0.mm.3D_firstorder_90Percentile	0.13	1	0.32
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	-280	0.35	0.55
wavelet.HHH_firstorder_RootMeanSquared	1800	0.36	0.55
wavelet.LLL_glcm_ClusterProminence	-0.026	0.31	0.58
wavelet.HHH_glcm_ClusterProminence	-280	0.21	0.65
original_glcm_Imc1	-130	0.15	0.69
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	910	0.03	0.87
wavelet.HLH_glcm_DifferenceAverage	-24	0.01	0.93

Table 5.10: Table showing results of univariate cox models fitted for each individual radiomic feature including b coefficient, Wald test result and p value. Table-wise Bonferroni corrected $\alpha = 0.0016$.

5.2.3 Clustering

On table 5.11 we can see the dataset used as input to fit the model based clustering, after performing min-max scaling of the variables.

described_variables	mean	sd	p00	p50	p100	skewness	kurtosis
age	0.5931502	0.1722010	0	0.5969193	1	-0.3111225	-0.2225480
original_shape_VoxelVolume	0.5829356	0.2252327	0	0.6175421	1	-0.4374695	-0.5534142
wavelet.HHL_glcm_ClusterProminence	0.0989860	0.1163366	0	0.0615282	1	2.7786461	12.3225476
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	0.2387529	0.2103486	0	0.2024770	1	0.9858591	0.5654748
log.sigma.2.0.mm.3D_glcm_Autocorrelation	0.2029748	0.1263963	0	0.1838236	1	2.3153056	8.8514924
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	0.0508914	0.0877837	0	0.0302539	1	6.2615524	50.2451503
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	0.5485125	0.2159559	0	0.5862543	1	-0.5484747	-0.3205776
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	0.0475191	0.0658836	0	0.0270760	1	8.4503877	115.9378509
wavelet.HHH_glszm_GrayLevelNonUniformity	0.0561877	0.1065138	0	0.0245006	1	5.0057489	31.3398745
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	0.3153470	0.1592984	0	0.2936679	1	0.9604226	1.5132056
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	0.0630528	0.0949388	0	0.0329395	1	4.8891785	35.0596521
wavelet.HHH_glcm_ClusterProminence	0.0755803	0.0663751	0	0.0659739	1	7.6866405	100.5859224
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	0.2740752	0.1561735	0	0.2616347	1	1.1819527	3.1492263
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	0.5865928	0.2011479	0	0.6240089	1	-0.6022647	-0.1985876
wavelet.LLL_glcm_ClusterProminence	0.5862742	0.1577118	0	0.6066546	1	-0.5711949	0.8381037
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	0.0998352	0.0973916	0	0.0696657	1	3.4677774	22.4587047
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	0.1630327	0.1725658	0	0.0976298	1	2.3361783	6.1841269
wavelet.HHH_glrlm_GrayLevelVariance	0.0673688	0.1069052	0	0.0246964	1	4.8122790	29.9362933
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	0.4370991	0.1545526	0	0.4409033	1	0.0031770	0.1886998
wavelet.HHH_glcm_Imc1	0.7284170	0.1749918	0	0.7690662	1	-1.0817978	1.2107301
original_glcm_Imc1	0.6018814	0.1204200	0	0.6094712	1	-0.4533090	1.5221494
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	0.3233587	0.1855941	0	0.2940825	1	0.6620482	0.3129117
wavelet.LHH_glszm_GrayLevelNonUniformity	0.1880146	0.1677015	0	0.1479853	1	1.4847299	2.6080324
wavelet.LHH_firstorder_Skewness	0.4731001	0.0843611	0	0.4633876	1	1.0663667	8.8730064
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	0.4335451	0.2156077	0	0.5136202	1	-0.0653329	-0.7465740
wavelet.LHH_glcm_Imc1	0.6823681	0.1463222	0	0.6952481	1	-1.0972745	2.1929685
log.sigma.5.0.mm.3D_firstorder_Kurtosis	0.1050651	0.1035381	0	0.0753691	1	3.5115187	20.3689620
log.sigma.5.0.mm.3D_firstorder_90Percentile	0.8477223	0.0711705	0	0.8576667	1	-5.4629759	56.1573713
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	0.2129439	0.1804696	0	0.1778410	1	2.1247949	4.5987508
wavelet.HLH_glcm_DifferenceAverage	0.2566230	0.1283222	0	0.2391226	1	1.7761862	5.4551338
wavelet.HHH_firstorder_RootMeanSquared	0.7815879	0.1032906	0	0.8015007	1	-3.6491017	19.5451207
wavelet.LHL_gldm_DependenceVariance	0.3538225	0.2104678	0	0.3097211	1	0.5354665	-0.3600384

Table 5.11: Dataset used as input to fit the model based clustering. Even though log transformation was applied as a first transformation, some variables are still quite asymmetric. Now all the numeric variables have the same min-max range of values.

In both models both models 10 was the number of components selected according to BIC criterion. In the model with variable selection 28 (84.85 %)) of the variables were considered relevant for clustering, 5 variables were excluded:

- age,
- gender,
- log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized,
- original_glcm_Imc1,
- wavelet.HHH_glcm_Imc1.

So the model including variable selection basically excluded the 2 clinical variables available on the final dataset and three radiomic variables.

Within both models three variables showed the highest discriminative power including: original_Shape_VoxelVolume (a 3D shape feature obtained from the original image, that measures the volume of the ROI by multiplying the number of voxels in the ROI by the volume of a single voxel), wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis and wavelet.LHL_glszm _LargeAreaHighGrayLevelEmphasis (Texture Gray Level Size Zone Matrix Features that measure the proportion of the joint distribution of larger size zones with higher gray-level values in the ROI, obtained from wavelet filtered images with different combinations of high-high-high and low-high-low pass filtering in the different image planes. In figure 5.16 and 5.17 we can see discriminative power calculated for each variable within each model. When representing most discriminative variables with empirical values against the theoretical fitted distribution within each model we can also see a good fit (Figure 5.18 a-f). Probabilities of misclassification were near to 0 for every cluster in both models, we can see these results represented with corresponding barplots for each cluster, in figure 5.19. An adjusted Rand index of 0.5 was obtained when comparing agreement between partitions generated by both models, so we can say there is some level of agreement, but two different models were in effect generated, they do not match perfectly.

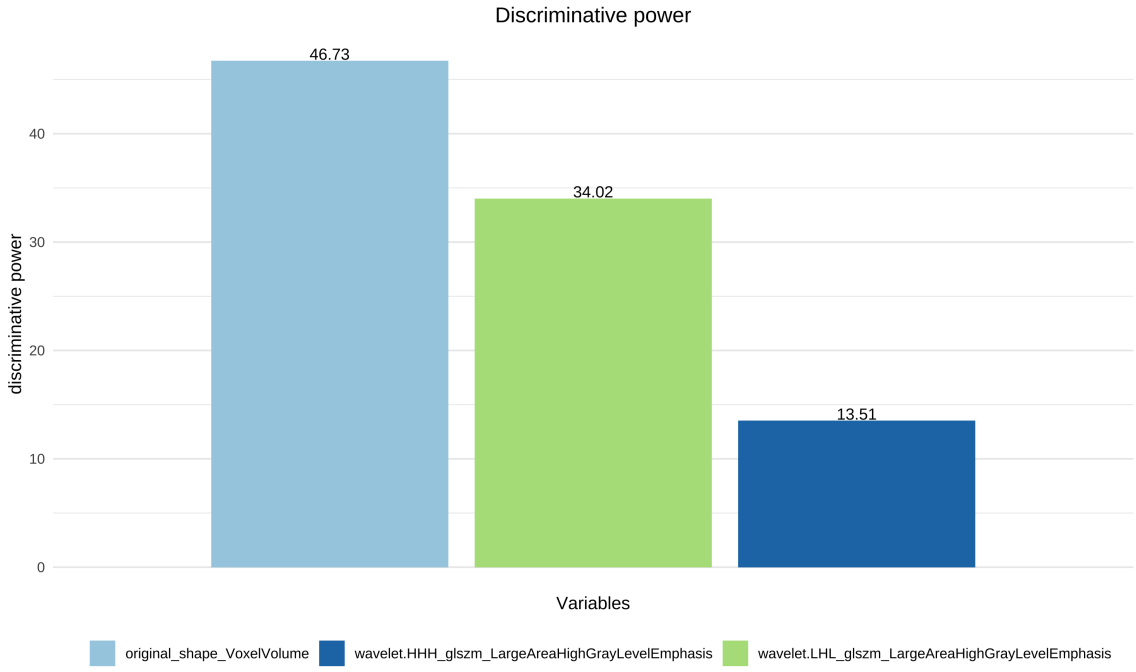


Figure 5.16: Barplot representing discrimination power for most discriminative variables within the model with wrapper variables selection.

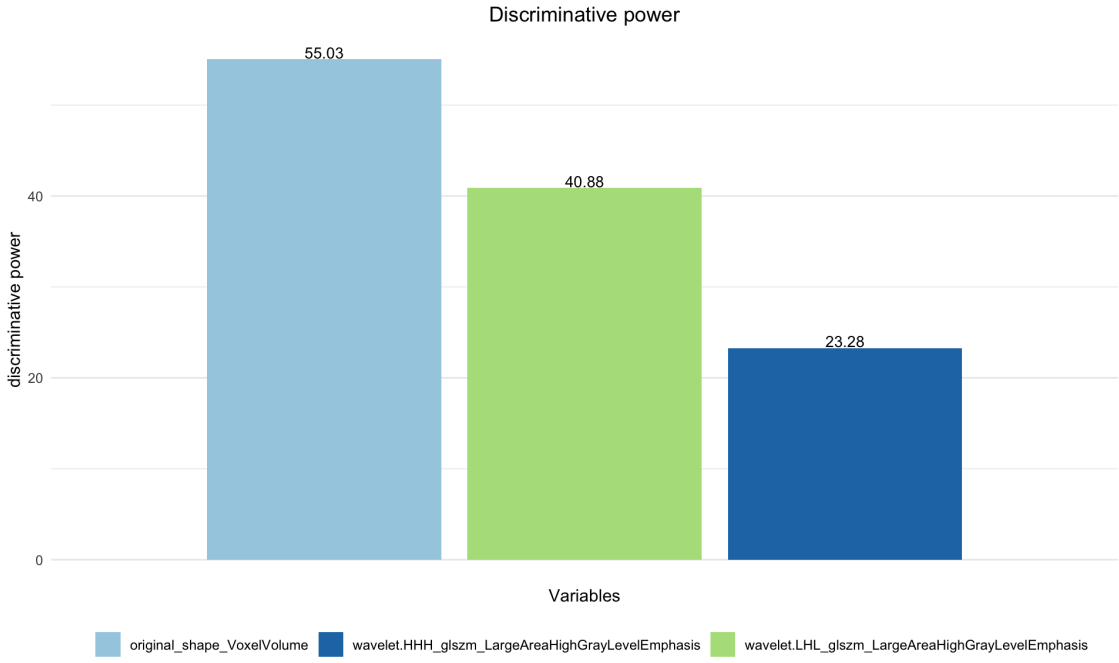


Figure 5.17: Barplot representing discrimination power for most discriminative variables within the model without wrapper variables selection.

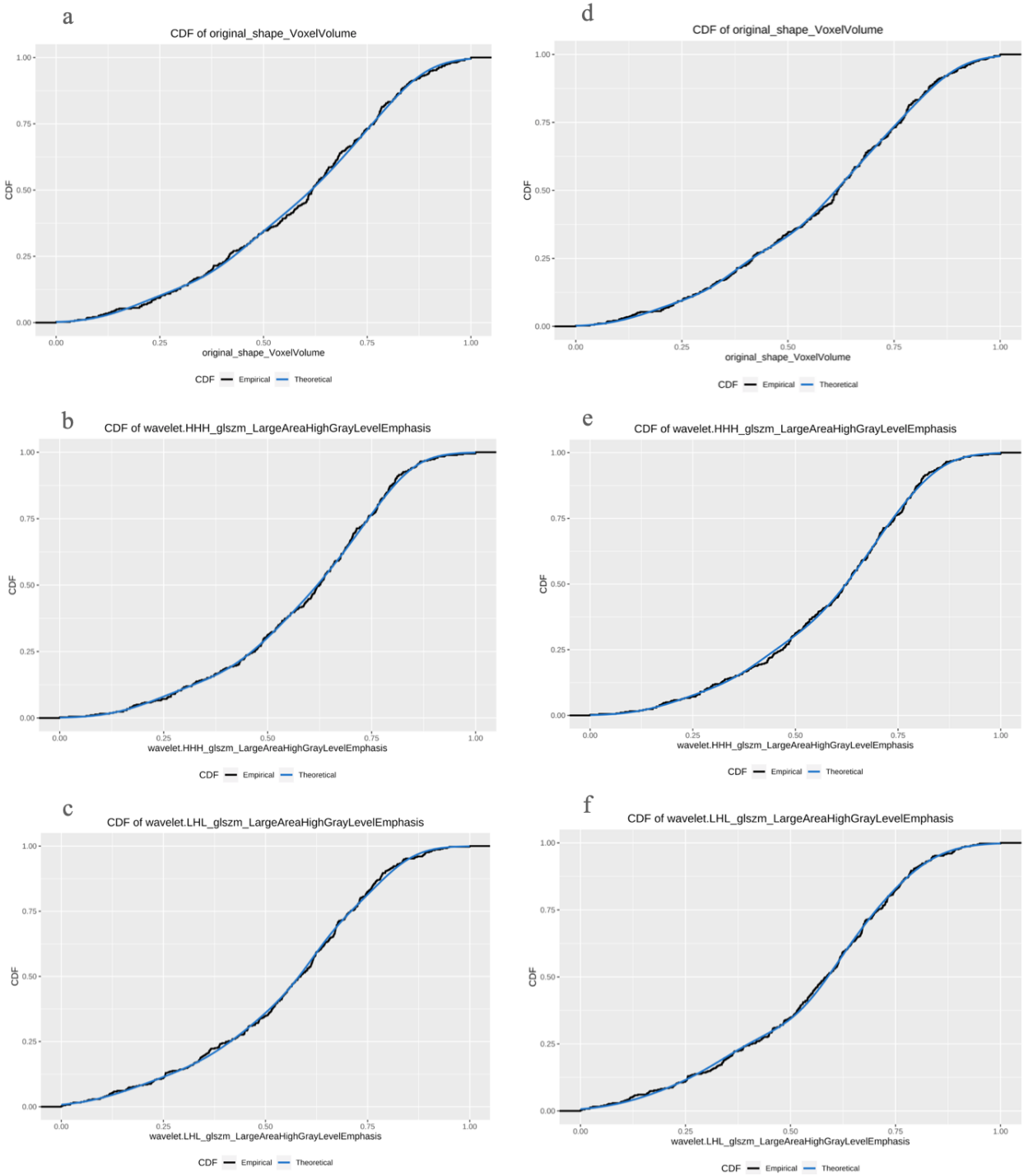


Figure 5.18: Empirical vs. theoretical fitted distribution for each of the most discriminative variables within the model with wrapper variable selection (figures a,b,c) and within the model without wrapper variable selection(figures d,e,f). We can verify the goodness of fit for the most discriminative variables.

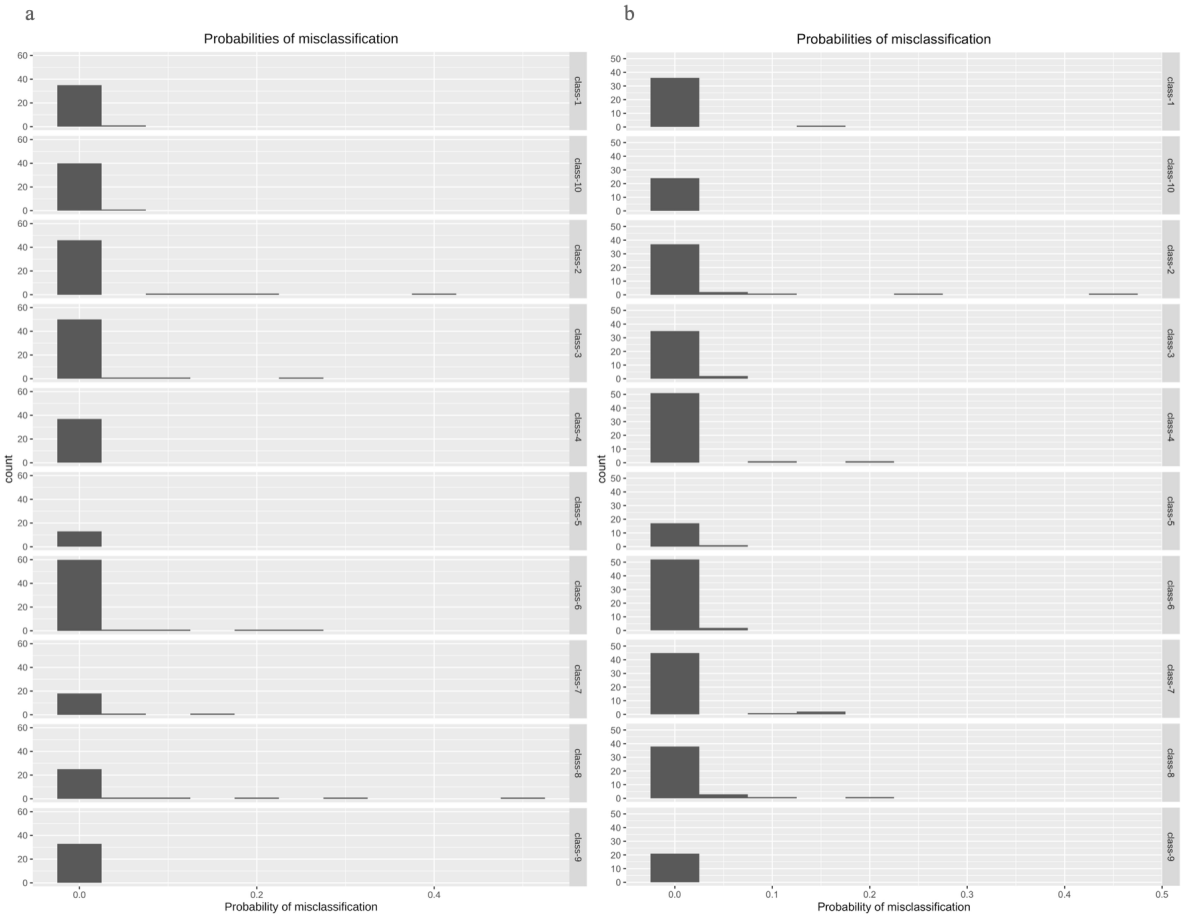


Figure 5.19: Barplots representing miss-classification probabilities for each cluster within the model with wrapper variable selection (figures a) and within the model without wrapper variable selection(figures b). We can see that the miss-classification probability is zero for most observations with some isolated exceptions.

When representing cluster groups using the first three components obtained after PCA analysis using numerical variables within the input data set, we can easily identify some common structure between partitions generated by both models and also some clear differences (Figure 5.20). For example cluster 6 from wrapper variable selection model (figure 5.20.a), seems partitioned between cluster 1 and 10 in the model without variable selection (figure 5.20.b). Then cluster 4 from the wrapper variable selection model (figure 5.20.a) does not appear as an independent cluster in the model without variable selection (figure 5.20.b), observations corresponding to this cluster in the first model seem to be divided amongst neighbor clusters in the second model.

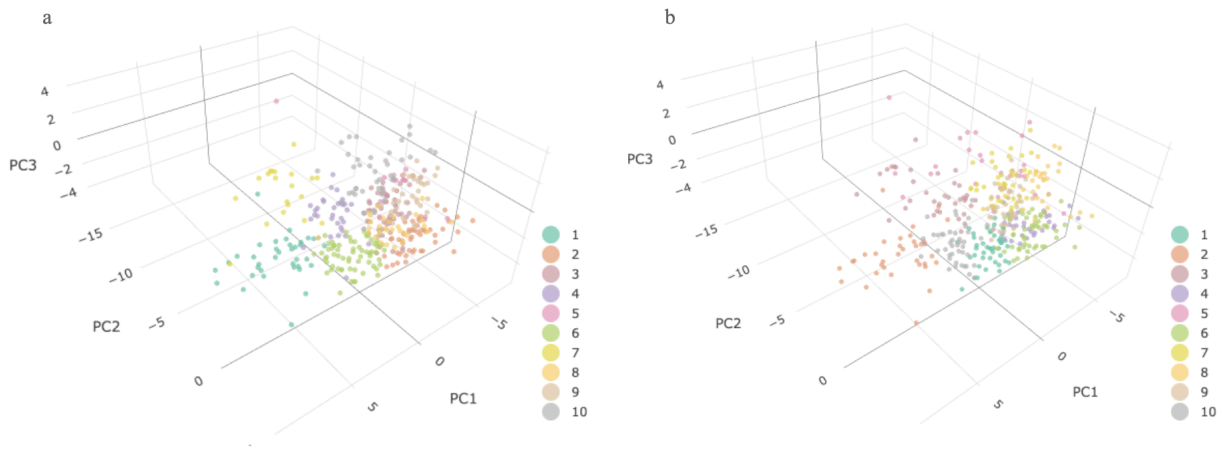


Figure 5.20: 3D Scatter plots representing each observation with its corresponding defined cluster color-coded and represented along the 3 principal components calculated obtained from PCA analysis from numeric input data set. Figure a corresponds to clusters generated by the model with wrapper variable selection, and figure b to the clusters generated by the model without wrapper variables selection.

	1	2	3	4	5	6	7	8	9	10	Sum
adenocarcinoma	0.019	0.019	0.013	0.019	0.013	0.011	0.008	0.011	0.011	0.013	0.135
large cell	0.034	0.016	0.042	0.029	0.019	0.037	0.042	0.024	0.032	0.027	0.302
nos	0.008	0.019	0.027	0.013	0.005	0.019	0.011	0.021	0.021	0.019	0.162
squamous cell carcinoma	0.056	0.037	0.016	0.024	0.037	0.066	0.056	0.021	0.034	0.053	0.401
Sum	0.117	0.090	0.098	0.085	0.074	0.133	0.117	0.077	0.098	0.111	1.000

Table 5.12: Relative joint and marginal frequencies between histology class and partitions generated by the model with wrapper variable selection.

	1	2	3	4	5	6	7	8	9	10	Sum
adenocarcinoma	0.021	0.016	0.011	0.013	0.008	0.027	0.003	0.013	0.008	0.016	0.135
large cell	0.016	0.034	0.042	0.037	0.003	0.048	0.019	0.029	0.024	0.050	0.302
nos	0.019	0.019	0.027	0.024	0.003	0.011	0.013	0.019	0.021	0.008	0.162
squamous cell carcinoma	0.040	0.064	0.061	0.024	0.021	0.085	0.019	0.019	0.034	0.034	0.401
Sum	0.095	0.133	0.141	0.098	0.034	0.170	0.053	0.080	0.088	0.109	1.000

Table 5.13: Relative joint and marginal frequencies between histology class and partitions generated by the model without wrapper variable selection.

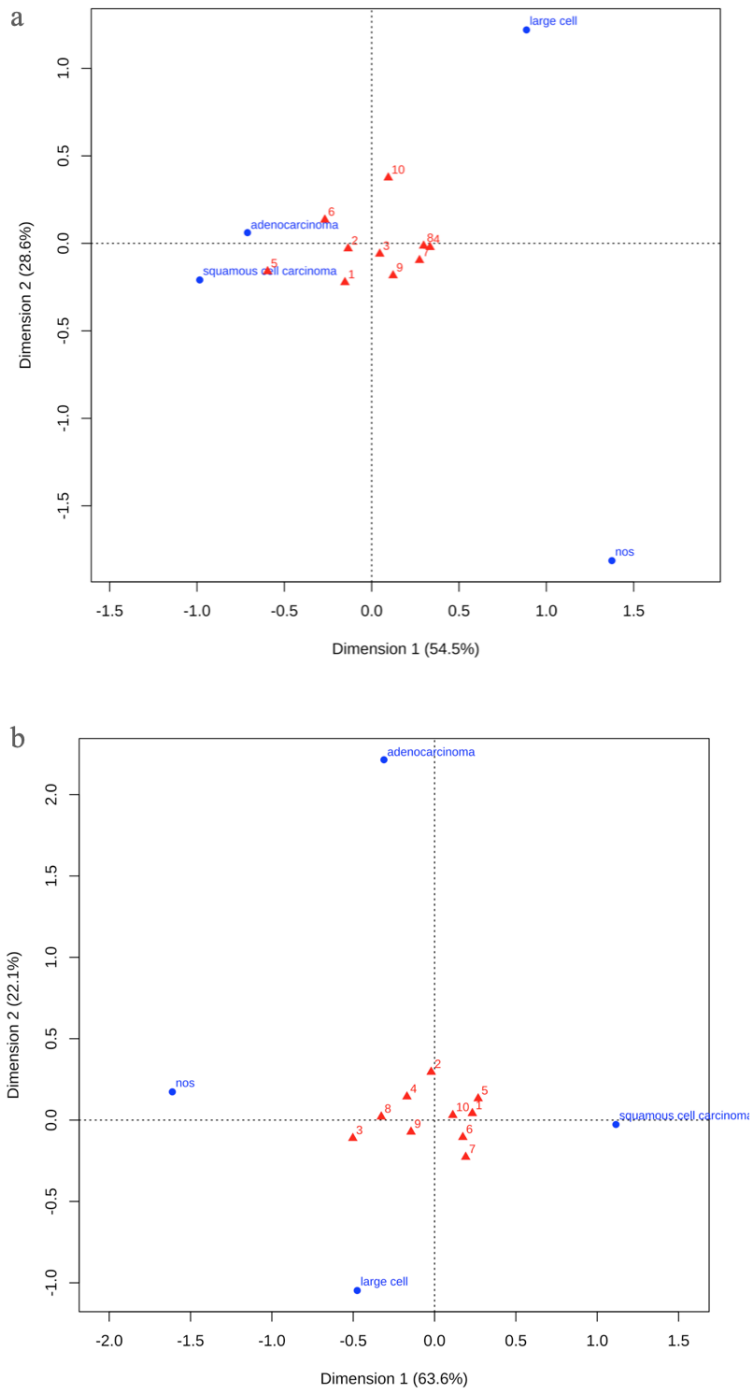


Figure 5.21: Correspondence analysis, asymmetric representation showing columns (clusters) represented with the principal coordinates and histology classes represented with standard coordinates. We can see associations between different clusters and between clusters and different histology classes. Clusters 6 and 10 showed the highest inertias for the model with wrapper variable selection (a). Clusters 3 and 7 showed highest inertias for the model without variable selection (b).

label	variable	6	10	pval	test
original_shape_VoxelVolume	Mean (std)	0.5 (0.1)	0.7 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHL_glcm_ClusterProminence	Mean (std)	0.03 (0.02)	0.2 (0.1)	<0.0001	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	0.1 (0.1)	0.3 (0.2)	<0.0001	(Wilcoxon rank sum test)
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	0.2 (0.1)	0.3 (0.1)	<0.0001	(Two Sample t-test)
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.5 (0.1)	0.6 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	0.01 (0.006)	0.1 (0.04)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	0.02 (0.01)	0.2 (0.2)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.4 (0.1)	<0.0001	(Two Sample t-test)
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.1 (0.04)	0.02 (0.01)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHH_glcm_ClusterProminence	Mean (std)	0.1 (0.03)	0.1 (0.04)	<0.0001	(Wilcoxon rank sum test)
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.3 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.5 (0.1)	0.6 (0.1)	<0.0001	(Two Sample t-test)
wavelet.LLL_glcm_ClusterProminence	Mean (std)	0.6 (0.1)	0.7 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.03 (0.01)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.1 (0.03)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	0.02 (1e-04)	0.2 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHH_glcm_Imc1	Mean (std)	0.8 (0.1)	0.7 (0.1)	<0.0001	(Wilcoxon rank sum test)
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.4 (0.1)	<0.0001	(Two Sample t-test)
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	0.1 (0.1)	0.4 (0.2)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	0.6 (0.1)	0.3 (0.1)	<0.0001	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	0.1 (0.04)	0.2 (0.1)	<0.0001	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	0.8 (0.1)	0.9 (0.03)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.3 (0.2)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	0.1 (0.04)	0.4 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	0.8 (0.05)	0.8 (0.04)	<0.0001	(Wilcoxon rank sum test)
Manufacturer	CMS Inc.	4 (21.05%)	15 (78.95%)	0.0001	(Pearson's Chi-squared test)
Manufacturer	SIEMENS	60 (69.77%)	26 (30.23%)	0.0001	(Pearson's Chi-squared test)
wavelet.LHL_gldm_DependenceVariance	Mean (std)	0.4 (0.2)	0.3 (0.2)	0.0002	(Wilcoxon rank sum test)
clinical.T.Stage	1	11 (91.67%)	1 (8.33%)	0.0017	(Fisher's Exact Test for Count Data)
clinical.T.Stage	2	35 (68.63%)	16 (31.37%)	0.0017	(Fisher's Exact Test for Count Data)
clinical.T.Stage	3	6 (66.67%)	3 (33.33%)	0.0017	(Fisher's Exact Test for Count Data)
clinical.T.Stage	4	12 (36.36%)	21 (63.64%)	0.0017	(Fisher's Exact Test for Count Data)
wavelet.LHH_glcm_Imc1	Mean (std)	0.6 (0.1)	0.7 (0.1)	0.0035	(Wilcoxon rank sum test)
wavelet.LHH_firstorder_Skewness	Mean (std)	0.5 (0.1)	0.5 (0.05)	0.0077	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.03 (0.02)	0.03 (0.02)	0.0292	(Wilcoxon rank sum test)
Overall.Stage	I	11 (78.57%)	3 (21.43%)	0.1191	(Fisher's Exact Test for Count Data)
Overall.Stage	II	4 (80.00%)	1 (20.00%)	0.1191	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIa	26 (66.67%)	13 (33.33%)	0.1191	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIb	23 (48.94%)	24 (51.06%)	0.1191	(Fisher's Exact Test for Count Data)
Histology	adenocarcinoma	10 (62.50%)	6 (37.50%)	0.2150	(Fisher's Exact Test for Count Data)
Histology	large cell	18 (48.65%)	19 (51.35%)	0.2150	(Fisher's Exact Test for Count Data)
Histology	nos	4 (57.14%)	3 (42.86%)	0.2150	(Fisher's Exact Test for Count Data)
Histology	squamous cell carcinoma	32 (71.11%)	13 (28.89%)	0.2150	(Fisher's Exact Test for Count Data)
age	Mean (std)	0.6 (0.2)	0.6 (0.1)	0.2274	(Two Sample t-test)
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	0.4 (0.1)	0.4 (0.2)	0.5356	(Two Sample t-test)
original_glcm_Imc1	Mean (std)	0.6 (0.1)	0.6 (0.1)	0.5677	(Wilcoxon rank sum test)
gender	female	22 (62.86%)	13 (37.14%)	0.7773	(Pearson's Chi-squared test)
gender	male	42 (60.00%)	28 (40.00%)	0.7773	(Pearson's Chi-squared test)
Clinical.N.Stage	0	15 (65.22%)	8 (34.78%)	0.8191	(Pearson's Chi-squared test)
Clinical.N.Stage	1_2	32 (58.18%)	23 (41.82%)	0.8191	(Pearson's Chi-squared test)
Clinical.N.Stage	3	17 (62.96%)	10 (37.04%)	0.8191	(Pearson's Chi-squared test)

Table 5.14: Table comparing distribution for different variables between clusters 6 and 10, from the model with wrapper variable selection, those showing highest inertia when performing correspondence analysis between partitions and histology class. Both variables included as input in the model or not are evaluated according to its distribution amongst selected clusters. Table-wise Bonferroni corrected $\alpha = 0.0013$.)

label	variable	3	7	pval	test
original_shape_VoxelVolume	Mean (std)	0.4 (0.1)	0.7 (0.1)	<0.0001	(Two Sample t-test)
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	0.05 (0.05)	0.2 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	0.2 (0.1)	0.3 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.3 (0.1)	0.7 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	0.02 (0.03)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.4 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.1 (0.05)	0.02 (0.007)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHH_glcm_ClusterProminence	Mean (std)	0.1 (0.05)	0.1 (0.02)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.3 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.4 (0.1)	0.7 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.1 (0.04)	0.03 (0.01)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.1 (0.03)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.4 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	0.1 (0.1)	0.4 (0.2)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	0.05 (0.03)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	0.8 (0.1)	0.9 (0.03)	<0.0001	(Welch Two Sample t-test)
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.1 (0.02)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHL_gldm_DependenceVariance	Mean (std)	0.2 (0.2)	0.4 (0.2)	0.0001	(Wilcoxon rank sum exact test)
clinical.T.Stage	1	15 (93.75%)	1 (6.25%)	0.0001	(Fisher's Exact Test for Count Data)
clinical.T.Stage	2	12 (37.50%)	20 (62.50%)	0.0001	(Fisher's Exact Test for Count Data)
clinical.T.Stage	3	3 (42.86%)	4 (57.14%)	0.0001	(Fisher's Exact Test for Count Data)
clinical.T.Stage	4	7 (26.92%)	19 (73.08%)	0.0001	(Fisher's Exact Test for Count Data)
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	0.7 (0.1)	0.8 (0.03)	0.0002	(Wilcoxon rank sum exact test)
wavelet.HHH_glcm_Imc1	Mean (std)	0.8 (0.1)	0.7 (0.1)	0.0018	(Wilcoxon rank sum exact test)
Histology	adenocarcinoma	5 (62.50%)	3 (37.50%)	0.0109	(Fisher's Exact Test for Count Data)
Histology	large cell	16 (50.00%)	16 (50.00%)	0.0109	(Fisher's Exact Test for Count Data)
Histology	nos	10 (71.43%)	4 (28.57%)	0.0109	(Fisher's Exact Test for Count Data)
Histology	squamous cell carcinoma	6 (22.22%)	21 (77.78%)	0.0109	(Fisher's Exact Test for Count Data)
wavelet.LHH_glcm_Imc1	Mean (std)	0.6 (0.2)	0.7 (0.1)	0.0123	(Wilcoxon rank sum exact test)
original_glcm_Imc1	Mean (std)	0.6 (0.1)	0.6 (0.1)	0.0336	(Wilcoxon rank sum exact test)
Overall.Stage	I	11 (78.57%)	3 (21.43%)	0.0402	(Fisher's Exact Test for Count Data)
Overall.Stage	II	1 (20.00%)	4 (80.00%)	0.0402	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIa	10 (40.00%)	15 (60.00%)	0.0402	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIb	15 (40.54%)	22 (59.46%)	0.0402	(Fisher's Exact Test for Count Data)
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	0.2 (0.2)	0.3 (0.1)	0.0501	(Wilcoxon rank sum test)
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	0.4 (0.2)	0.3 (0.1)	0.0622	(Wilcoxon rank sum exact test)
wavelet.LLL_glcm_ClusterProminence	Mean (std)	0.7 (0.1)	0.7 (0.1)	0.0814	(Two Sample t-test)
wavelet.HHL_glcm_ClusterProminence	Mean (std)	0.2 (0.1)	0.2 (0.1)	0.0925	(Wilcoxon rank sum exact test)
wavelet.LHH_firstorder_Skewness	Mean (std)	0.5 (0.1)	0.5 (0.05)	0.2208	(Wilcoxon rank sum exact test)
Clinical.N.Stage	0	15 (51.72%)	14 (48.28%)	0.2900	(Pearson's Chi-squared test)
Clinical.N.Stage	1_2	13 (36.11%)	23 (63.89%)	0.2900	(Pearson's Chi-squared test)
Clinical.N.Stage	3	9 (56.25%)	7 (43.75%)	0.2900	(Pearson's Chi-squared test)
gender	female	16 (51.61%)	15 (48.39%)	0.3986	(Pearson's Chi-squared test)
gender	male	21 (42.00%)	29 (58.00%)	0.3986	(Pearson's Chi-squared test)
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	0.5 (0.2)	0.5 (0.1)	0.5070	(Two Sample t-test)
age	Mean (std)	0.6 (0.2)	0.6 (0.2)	0.6372	(Two Sample t-test)
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	0.1 (0.1)	0.1 (0.03)	0.7307	(Wilcoxon rank sum exact test)
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	0.1 (0.05)	0.1 (0.03)	0.7665	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.04 (0.03)	0.04 (0.03)	0.7954	(Wilcoxon rank sum exact test)
Manufacturer	CMS Inc.	10 (43.48%)	13 (56.52%)	0.8023	(Pearson's Chi-squared test)
Manufacturer	SIEMENS	27 (46.55%)	31 (53.45%)	0.8023	(Pearson's Chi-squared test)

Table 5.15: Table comparing distribution for different variables between clusters 3 and 7, from the model without variable selection, those showing highest inertia when performing correspondence between partitions and histology class. Both variables included as input in the model or not are evaluated according to its distribution amongst selected clusters. Table-wise Bonferroni corrected $\alpha = 0.0013$.)

When evaluating the relationship between partitions and histology classes, ARI index was 0.02 for the model with variable selection, and 0.05 for the model without variable selection, mainly in favor of random labeling if we take into account all clusters. After correspondence analysis I found the two clusters showing the highest inertias were clusters 6 and 10 for the model with wrapper variable selection and clusters 3 and 7 for the model without variable selection

(Figure 5.21). This pair of clusters was further analyzed in both cases to evaluate distribution of different variables amongst both clusters and compare them (Tables 5.14 and 5.15). For the clusters selected for the first model (with variable selection) no significant relationship was found with histology class. Other variables did show significant differences between these clusters as clinical T stage and original_shape_VoxelVolume, Manufacturer model, and many texture radiomic variables applied over wavelet and LoG filtered images (Table 5.14). For the clusters selected for the second model (without variable selection) test for independence showed a p value of 0.0109 in the limit of significance, though not enough if we take multiple comparisons correction into account. Proportion of adenocarcinoma and nos histotypes was higher within cluster 3 than 7, and squamous cell carcinoma was mostly present within cluster 7, large cell cancer was equally distributed among both clusters. Interestingly, overall stage and age were not related to partitions generated by these clusters so these variables, previously found related to histology classes, should not condition these results significantly. Clinical T stage and original_shape_VoxelVolume were again significantly related to the clusters evaluated, with cluster 7 showing greater average volume and more presence within higher T stage categories. Manufacturer model or other clinical variables showed no significant relationship with these clusters. Main related radiomic features, after original_shape_VoxelVolume, included mainly texture features obtained from images with LoG and different wavelet filtering. Only isolated radiomic features matched those previously found as possibly related with histology variable when analyzed in an univariate manner, these included wavelet.HHH_gldm_ClusterProminence, wavelet.LHL_gldm_DependenceVariance and wavelet.HHH_glcm_lmc1 (Tables 5.8 and 5.14). wavelet.LHL_gldm_DependenceVariance showed a higher mean for cluster 7 vs cluster 3, and wavelet.HHH_glcm_lmc1 slightly lower for cluster 7 vs cluster 3. (Table 5.14)

A similar analysis was performed to evaluate association of different radiomic features to overall stage categories, clusters 2 and 4 showed the highest inertias for the model with wrapper variable selection, clusters 6 and 10 showed highest inertias for the model without variable selection (Figure 5.22). In both cases a significant relationship between overall stage and selected clusters was found, clinical T stage also showed a significant relationship and again original_shape_VoxelVolume led the table results (Tables 5.18 and 5.19). More advanced overall stages, clinical T stage and voxel volume, were present within cluster 2, while the opposite was true for cluster 4. As previously mentioned in both models this was the most discriminative variable, so it makes sense that clinical T stage shows a significant relationship with different combinations of clusters, and given its correlation to overall stage, it makes sense also to find these three variables significantly related together. Other main radiomic features heading the table were wavelet.HHl_glcm_ClusterProminence showing greater values within cluster 4 (earlier stages), and log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis both showing greater values within cluster 9 (more advanced stages). These results were similar for both models when comparing clusters with earlier and greatest stages.

	1	2	3	4	5	6	7	8	9	10	Sum
I	0.019	0.005	0.016	0.037	0.011	0.029	0.011	0.016	0.021	0.008	0.172
II	0.011	0.019	0.024	0.008	0.011	0.011	0.003	0.000	0.013	0.003	0.101
IIIa	0.034	0.027	0.042	0.027	0.005	0.069	0.016	0.021	0.011	0.034	0.286
IIIb	0.032	0.082	0.058	0.027	0.008	0.061	0.024	0.042	0.042	0.064	0.440
Sum	0.095	0.133	0.141	0.098	0.034	0.170	0.053	0.080	0.088	0.109	1.000

Table 5.16: Relative joint and marginal frequencies between Overall stage category and partitions generated by the model with wrapper variable selection.

	1	2	3	4	5	6	7	8	9	10	Sum
I	0.013	0.013	0.029	0.008	0.013	0.005	0.008	0.019	0.024	0.040	0.172
II	0.011	0.013	0.003	0.005	0.013	0.016	0.011	0.008	0.019	0.003	0.101
IIIa	0.053	0.032	0.027	0.024	0.019	0.019	0.040	0.011	0.029	0.034	0.286
IIIb	0.040	0.032	0.040	0.048	0.029	0.093	0.058	0.040	0.027	0.034	0.440
Sum	0.117	0.090	0.098	0.085	0.074	0.133	0.117	0.077	0.098	0.111	1.000

Table 5.17: Relative joint and marginal frequencies between Overall stage category and partitions generated by the model without wrapper variable selection.

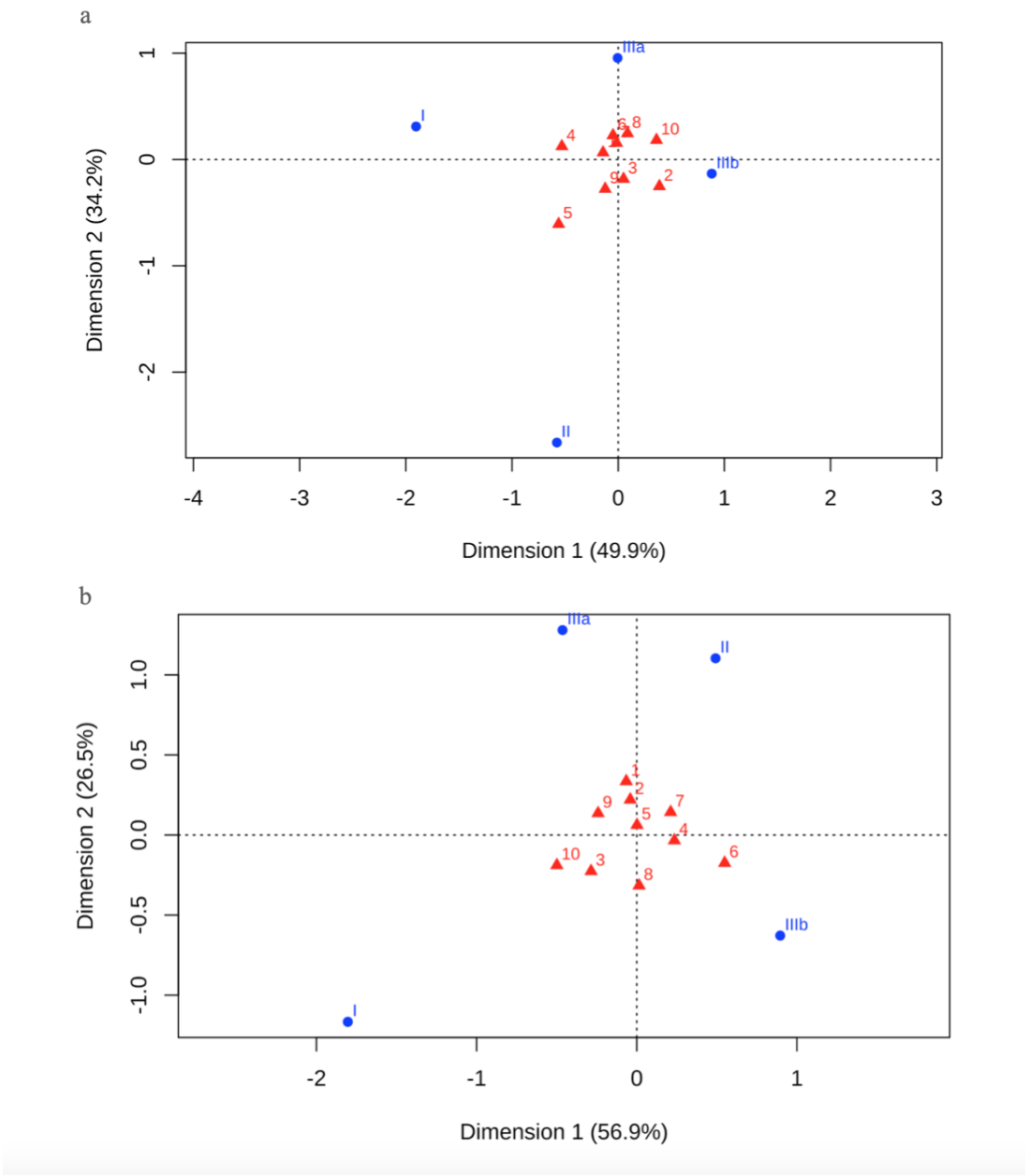


Figure 5.22: Correspondence analysis, asymmetric representation showing columns (clusters) represented with the principal coordinates and overall stage categories represented with standard coordinates. We can see associations between different clusters and between clusters and different stage categories . Clusters 2 and 4 showed the highest inertias for the model with wrapper variable selection (a). Clusters 6 and 10 showed highest inertias for the model without variable selection (b).

label	variable	2	4	pval	test
original_shape_VoxelVolume	Mean (std)	0.8 (0.1)	0.4 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHL_glcm_ClusterProminence	Mean (std)	0.02 (0.01)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	0.5 (0.2)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.8 (0.1)	0.4 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	0.01 (0.006)	0.1 (0.03)	<0.0001	(Welch Two Sample t-test)
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	0.04 (0.03)	0.01 (0.01)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.4 (0.1)	0.2 (0.1)	<0.0001	(Two Sample t-test)
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.02 (0.01)	0.1 (0.04)	<0.0001	(Wilcoxon rank sum test)
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.4 (0.1)	0.2 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.8 (0.1)	0.5 (0.1)	<0.0001	(Two Sample t-test)
wavelet.LLL_glcm_ClusterProminence	Mean (std)	0.5 (0.2)	0.6 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.1 (0.02)	0.2 (0.1)	<0.0001	(Wilcoxon rank sum test)
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.5 (0.1)	0.2 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	0.2 (0.1)	0.1 (0.05)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	0.6 (0.1)	0.3 (0.2)	<0.0001	(Wilcoxon rank sum test)
wavelet.LHH_glcm_Imc1	Mean (std)	0.8 (0.1)	0.7 (0.1)	<0.0001	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	0.1 (0.1)	0.05 (0.03)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	0.2 (0.04)	0.3 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.LHL_gldm_DependenceVariance	Mean (std)	0.5 (0.2)	0.3 (0.2)	<0.0001	(Wilcoxon rank sum test)
clinical.T.Stage	1	0 (0%)	15 (100.00%)	<0.0001	(Pearson's Chi-squared test)
clinical.T.Stage	2	14 (46.67%)	16 (53.33%)	<0.0001	(Pearson's Chi-squared test)
clinical.T.Stage	3	13 (81.25%)	3 (18.75%)	<0.0001	(Pearson's Chi-squared test)
clinical.T.Stage	4	23 (88.46%)	3 (11.54%)	<0.0001	(Pearson's Chi-squared test)
Overall.Stage	I	2 (12.50%)	14 (87.50%)	0.0001	(Fisher's Exact Test for Count Data)
Overall.Stage	II	7 (70.00%)	3 (30.00%)	0.0001	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIa	10 (50.00%)	10 (50.00%)	0.0001	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIb	31 (75.61%)	10 (24.39%)	0.0001	(Fisher's Exact Test for Count Data)
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	0.02 (4e-04)	0.1 (0.05)	0.0011	(Wilcoxon rank sum test)
original_glcm_Imc1	Mean (std)	0.6 (0.1)	0.7 (0.1)	0.0016	(Two Sample t-test)
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	0.9 (0.01)	0.8 (0.1)	0.0071	(Welch Two Sample t-test)
gender	female	13 (39.39%)	20 (60.61%)	0.0077	(Pearson's Chi-squared test)
gender	male	37 (68.52%)	17 (31.48%)	0.0077	(Pearson's Chi-squared test)
wavelet.LHH_firstorder_Skewness	Mean (std)	0.5 (0.1)	0.5 (0.1)	0.0134	(Wilcoxon rank sum test)
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	0.2 (0.1)	0.2 (0.1)	0.0226	(Two Sample t-test)
Manufacturer	CMS Inc.	5 (35.71%)	9 (64.29%)	0.0722	(Pearson's Chi-squared test)
Manufacturer	SIEMENS	45 (61.64%)	28 (38.36%)	0.0722	(Pearson's Chi-squared test)
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.1 (0.04)	0.1 (0.03)	0.1084	(Wilcoxon rank sum test)
wavelet.HHH_glcm_Imc1	Mean (std)	0.8 (0.1)	0.8 (0.2)	0.1084	(Wilcoxon rank sum test)
age	Mean (std)	0.6 (0.2)	0.6 (0.2)	0.1103	(Wilcoxon rank sum test)
Histology	adenocarcinoma	6 (54.55%)	5 (45.45%)	0.1386	(Fisher's Exact Test for Count Data)
Histology	large cell	13 (48.15%)	14 (51.85%)	0.1386	(Fisher's Exact Test for Count Data)
Histology	nos	7 (43.75%)	9 (56.25%)	0.1386	(Fisher's Exact Test for Count Data)
Histology	squamous cell carcinoma	24 (72.73%)	9 (27.27%)	0.1386	(Fisher's Exact Test for Count Data)
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.2 (0.2)	0.1421	(Wilcoxon rank sum test)
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	0.8 (0.02)	0.8 (0.1)	0.2327	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.03 (0.02)	0.1 (0.1)	0.2871	(Wilcoxon rank sum test)
wavelet.HHH_glcm_ClusterProminence	Mean (std)	0.1 (0.03)	0.1 (0.03)	0.3235	(Wilcoxon rank sum test)
Clinical.N.Stage	0	16 (48.48%)	17 (51.52%)	0.4154	(Pearson's Chi-squared test)
Clinical.N.Stage	1_2	22 (62.86%)	13 (37.14%)	0.4154	(Pearson's Chi-squared test)
Clinical.N.Stage	3	12 (63.16%)	7 (36.84%)	0.4154	(Pearson's Chi-squared test)
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	0.5 (0.1)	0.5 (0.2)	0.7863	(Two Sample t-test)

Table 5.18: Table comparing distribution for different variables between clusters 2 and 4, from the model without variable selection, those showing highest inertia when performing correspondence analysis between partitions and overall stage category. Both variables included as input in the model or not are evaluated according to its distribution amongst selected clusters. Table-wise Bonferroni corrected $\alpha = 0.0013$.)

label	variable	6	10	pval	test
original_shape_VoxelVolume	Mean (std)	0.8 (0.1)	0.4 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHL_glcm_ClusterProminence	Mean (std)	0.02 (0.02)	0.1 (0.04)	<0.0001	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	0.4 (0.2)	0.1 (0.05)	<0.0001	(Wilcoxon rank sum test)
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	0.2 (0.1)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.8 (0.1)	0.4 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	0.04 (0.03)	0.01 (0.01)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.4 (0.1)	0.2 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.02 (0.009)	0.1 (0.04)	<0.0001	(Welch Two Sample t-test)
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.4 (0.1)	0.1 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.8 (0.1)	0.4 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.1 (0.03)	0.2 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.1 (0.03)	0.3 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	0.03 (0.002)	0.02 (4e-04)	<0.0001	(Wilcoxon rank sum test)
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.5 (0.1)	0.1 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	0.2 (0.1)	0.1 (0.03)	<0.0001	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	0.1 (0.1)	0.05 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.3 (0.2)	<0.0001	(Wilcoxon rank sum test)
wavelet.LHL_gldm_DependenceVariance	Mean (std)	0.5 (0.2)	0.3 (0.2)	<0.0001	(Wilcoxon rank sum test)
Overall.Stage	I	2 (11.76%)	15 (88.24%)	<0.0001	(Fisher's Exact Test for Count Data)
Overall.Stage	II	6 (85.71%)	1 (14.29%)	<0.0001	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIa	7 (35.00%)	13 (65.00%)	<0.0001	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIb	35 (72.92%)	13 (27.08%)	<0.0001	(Fisher's Exact Test for Count Data)
clinical.T.Stage	1	0 (0%)	19 (100.00%)	<0.0001	(Pearson's Chi-squared test)
clinical.T.Stage	2	12 (42.86%)	16 (57.14%)	<0.0001	(Pearson's Chi-squared test)
clinical.T.Stage	3	8 (72.73%)	3 (27.27%)	<0.0001	(Pearson's Chi-squared test)
clinical.T.Stage	4	30 (88.24%)	4 (11.76%)	<0.0001	(Pearson's Chi-squared test)
wavelet.LHH_glcm_Imc1	Mean (std)	0.7 (0.1)	0.6 (0.1)	0.0009	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	0.9 (0.02)	0.8 (0.1)	0.0018	(Wilcoxon rank sum test)
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	0.01 (0.007)	0.02 (0.01)	0.0113	(Wilcoxon rank sum test)
wavelet.HHH_glcm_ClusterProminence	Mean (std)	0.1 (0.02)	0.1 (0.04)	0.0183	(Welch Two Sample t-test)
original_glcm_Imc1	Mean (std)	0.6 (0.1)	0.6 (0.1)	0.0203	(Wilcoxon rank sum test)
gender	female	12 (40.00%)	18 (60.00%)	0.0546	(Pearson's Chi-squared test)
gender	male	38 (61.29%)	24 (38.71%)	0.0546	(Pearson's Chi-squared test)
wavelet.LLL_glcm_ClusterProminence	Mean (std)	0.5 (0.2)	0.6 (0.1)	0.0548	(Wilcoxon rank sum test)
wavelet.LHH_firstorder_Skewness	Mean (std)	0.5 (0.1)	0.5 (0.1)	0.1264	(Wilcoxon rank sum test)
wavelet.HHH_glcm_Imc1	Mean (std)	0.8 (0.1)	0.8 (0.1)	0.1283	(Wilcoxon rank sum test)
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	0.6 (0.2)	0.6 (0.1)	0.4420	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.03 (0.02)	0.04 (0.1)	0.5255	(Wilcoxon rank sum test)
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	0.5 (0.1)	0.4 (0.2)	0.5832	(Wilcoxon rank sum test)
age	Mean (std)	0.5 (0.2)	0.6 (0.2)	0.5985	(Two Sample t-test)
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	0.8 (0.03)	0.8 (0.1)	0.7718	(Wilcoxon rank sum test)
Clinical.N.Stage	0	20 (57.14%)	15 (42.86%)	0.8764	(Pearson's Chi-squared test)
Clinical.N.Stage	1_2	20 (54.05%)	17 (45.95%)	0.8764	(Pearson's Chi-squared test)
Clinical.N.Stage	3	10 (50.00%)	10 (50.00%)	0.8764	(Pearson's Chi-squared test)
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	0.2 (0.03)	0.2 (0.1)	0.8907	(Welch Two Sample t-test)
Histology	adenocarcinoma	4 (44.44%)	5 (55.56%)	0.9043	(Fisher's Exact Test for Count Data)
Histology	large cell	14 (58.33%)	10 (41.67%)	0.9043	(Fisher's Exact Test for Count Data)
Histology	nos	7 (50.00%)	7 (50.00%)	0.9043	(Fisher's Exact Test for Count Data)
Histology	squamous cell carcinoma	25 (55.56%)	20 (44.44%)	0.9043	(Fisher's Exact Test for Count Data)
Manufacturer	CMS Inc.	5 (50.00%)	5 (50.00%)	1.0000	(Fisher's Exact Test for Count Data)
Manufacturer	SIEMENS	45 (54.88%)	37 (45.12%)	1.0000	(Fisher's Exact Test for Count Data)

Table 5.19: Table comparing distribution for different variables between clusters 6 and 10, from the model without variable selection, those showing highest inertia when performing correspondence analysis between partitions and overall stage category. Both variables included as input in the model or not are evaluated according to its distribution amongst selected clusters. Table-wise Bonferroni corrected $\alpha = 0.0013$.)

Regarding relationship between survival and partitions generated by the different models, a most significant difference was found between survival curves generated for clusters generated by the first model (with variable selection). The model without variable selection also showed some degree of relationship though in the limit of significance if we take into account correction for multiple comparisons. For the model generated with variable selection, clusters 7 and 9

showed the maximum and minimum median survival, for the model generated without variable selection clusters 3 and 8 showed the maximum and minimum median survival (Figure 5.23). This pair of clusters was further analyzed in both cases to evaluate distribution of different variables amongst clusters and compare them (Tables 5.20 and 5.21). Clinical T stage and original_shape_VoxelVolume were again significantly related to the clusters evaluated, with cluster 9 (the one with minimum median survival) showing greater average volume and more presence within higher T stage categories. Clinical T1 category was 100% within cluster 7. Other main radiomic features heading the table were wavelet.HHl_glcm_ClusterProminence showing greater values within cluster 7 (greater median survival), and

log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and

wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis both showing greater values within cluster 9 (minimum median survival). These results were similar for both models when comparing clusters with minimum and maximum median survival. There was no significant relationship between the manufacturer model and these clusters in either of both models.

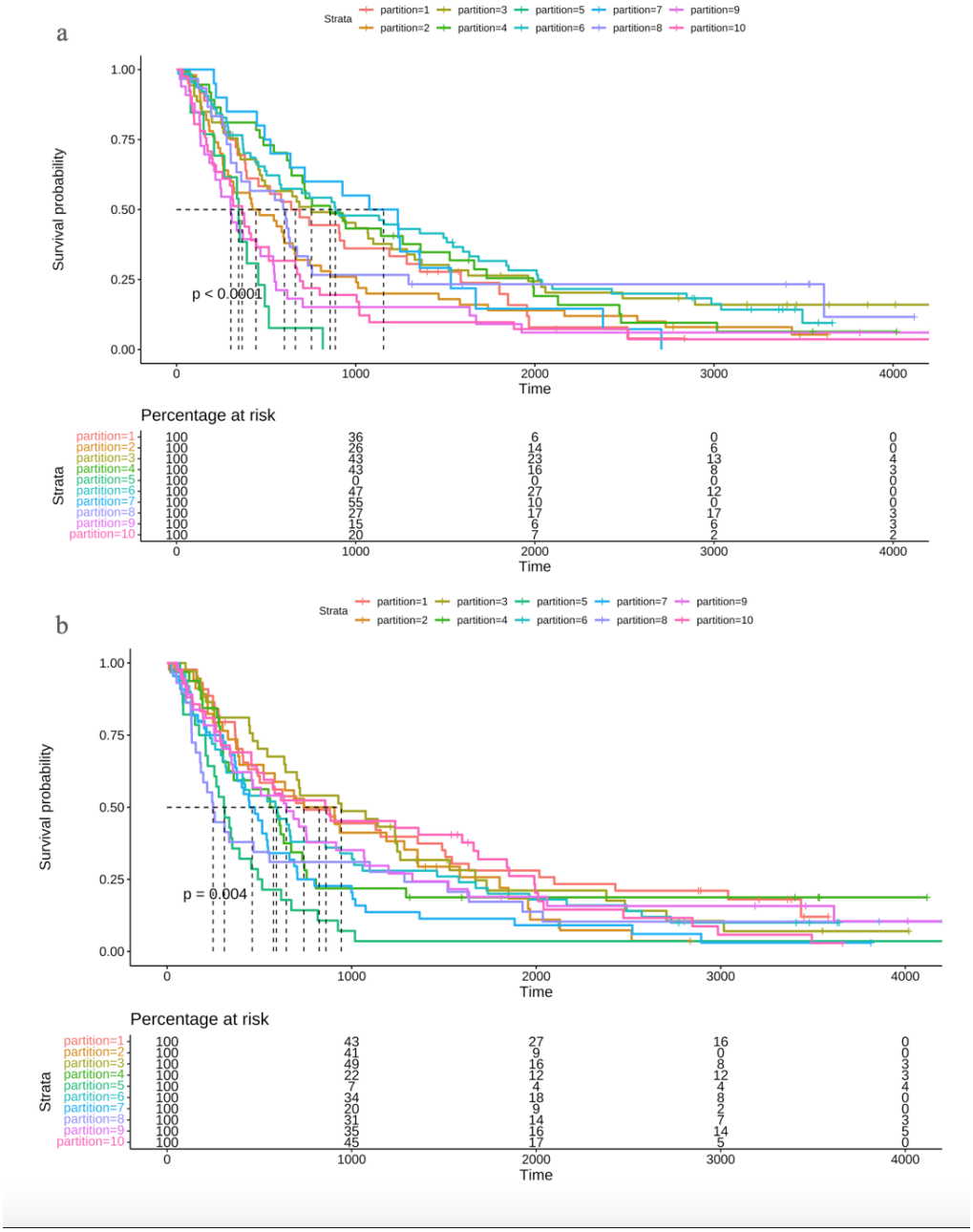


Figure 5.23: Kaplan Meyer plot showing survival probability over time in days for patients corresponding to different clusters for the model with wrapper variable selection (a) and the model without variable selection (b). Median survival for each group is indicated with the dotted line and p value corresponding to the log-rank test between curves is indicated on the plot. On the bottom we can see the percentage of patients at risk at each selected time point for each group. We can see a more significant difference between curves of clusters generated with model a. For the model generated with variable selection, clusters 7 and 9 showed the maximum and minimum median survival, for the model generated without variable selection clusters 3 and 8 showed the maximum and minimum median survival.

label	variable	7	9	pval	test
original_shape_VoxelVolume	Mean (std)	0.3 (0.1)	0.8 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHL_glcm_ClusterProminence	Mean (std)	0.3 (0.2)	0.1 (0.04)	<0.0001	(Welch Two Sample t-test)
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	0.03 (0.1)	0.4 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.2 (0.2)	0.8 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	0.03 (0.04)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.3 (0.1)	0.5 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.1 (0.1)	0.01 (0.006)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHH_glcm_ClusterProminence	Mean (std)	0.1 (0.1)	0.04 (0.03)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.4 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.3 (0.1)	0.8 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.1 (0.1)	0.03 (0.02)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.3 (0.2)	0.04 (0.02)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.5 (0.1)	<0.0001	(Two Sample t-test)
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	0.1 (0.04)	0.4 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	0.04 (0.04)	0.2 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.1 (0.02)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	0.5 (0.2)	0.3 (0.04)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHL_gldm_DependenceVariance	Mean (std)	0.2 (0.2)	0.4 (0.2)	<0.0001	(Wilcoxon rank sum exact test)
clinical.T.Stage	1	9 (100.00%)	0 (0%)	<0.0001	(Fisher's Exact Test for Count Data)
clinical.T.Stage	2	2 (10.53%)	17 (89.47%)	<0.0001	(Fisher's Exact Test for Count Data)
clinical.T.Stage	3	4 (44.44%)	5 (55.56%)	<0.0001	(Fisher's Exact Test for Count Data)
clinical.T.Stage	4	5 (31.25%)	11 (68.75%)	<0.0001	(Fisher's Exact Test for Count Data)
wavelet.LHH_glcm_Imc1	Mean (std)	0.5 (0.2)	0.7 (0.1)	0.0001	(Welch Two Sample t-test)
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.1 (0.03)	0.03 (0.03)	0.0003	(Wilcoxon rank sum exact test)
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	0.1 (0.1)	0.1 (0.03)	0.0005	(Welch Two Sample t-test)
wavelet.HHH_glcm_Imc1	Mean (std)	0.7 (0.2)	0.6 (0.2)	0.0029	(Wilcoxon rank sum exact test)
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	0.3 (0.3)	0.1 (0.02)	0.0037	(Wilcoxon rank sum exact test)
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	0.2 (0.1)	0.3 (0.1)	0.0571	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	0.8 (0.2)	0.9 (0.03)	0.0796	(Wilcoxon rank sum exact test)
wavelet.LLL_glcm_ClusterProminence	Mean (std)	0.7 (0.2)	0.7 (0.1)	0.1009	(Wilcoxon rank sum exact test)
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	0.3 (0.2)	0.2 (0.1)	0.1145	(Wilcoxon rank sum test)
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	0.7 (0.2)	0.8 (0.03)	0.1457	(Wilcoxon rank sum exact test)
Clinical.N.Stage	0	7 (26.92%)	19 (73.08%)	0.2409	(Fisher's Exact Test for Count Data)
Clinical.N.Stage	1_2	7 (43.75%)	9 (56.25%)	0.2409	(Fisher's Exact Test for Count Data)
Clinical.N.Stage	3	6 (54.55%)	5 (45.45%)	0.2409	(Fisher's Exact Test for Count Data)
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	0.5 (0.2)	0.4 (0.2)	0.2533	(Two Sample t-test)
Manufacturer	CMS Inc.	6 (30.00%)	14 (70.00%)	0.3657	(Pearson's Chi-squared test)
Manufacturer	SIEMENS	14 (42.42%)	19 (57.58%)	0.3657	(Pearson's Chi-squared test)
wavelet.LHH_firstorder_Skewness	Mean (std)	0.5 (0.2)	0.5 (0.05)	0.3768	(Wilcoxon rank sum exact test)
Overall.Stage	I	4 (33.33%)	8 (66.67%)	0.3921	(Fisher's Exact Test for Count Data)
Overall.Stage	II	1 (16.67%)	5 (83.33%)	0.3921	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIa	6 (60.00%)	4 (40.00%)	0.3921	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIb	9 (36.00%)	16 (64.00%)	0.3921	(Fisher's Exact Test for Count Data)
original_glcm_Imc1	Mean (std)	0.6 (0.1)	0.6 (0.1)	0.4544	(Two Sample t-test)
age	Mean (std)	0.6 (0.2)	0.6 (0.2)	0.4649	(Two Sample t-test)
gender	female	7 (43.75%)	9 (56.25%)	0.5525	(Pearson's Chi-squared test)
gender	male	13 (35.14%)	24 (64.86%)	0.5525	(Pearson's Chi-squared test)
Histology	adenocarcinoma	1 (25.00%)	3 (75.00%)	0.9506	(Fisher's Exact Test for Count Data)
Histology	large cell	7 (43.75%)	9 (56.25%)	0.9506	(Fisher's Exact Test for Count Data)
Histology	nos	5 (38.46%)	8 (61.54%)	0.9506	(Fisher's Exact Test for Count Data)
Histology	squamous cell carcinoma	7 (35.00%)	13 (65.00%)	0.9506	(Fisher's Exact Test for Count Data)

Table 5.20: Table comparing distribution for different variables between clusters 7 and 9, from the model without variable selection, those showing greatest difference in median survival. Both variables included as input in the model or not are evaluated according to its distribution amongst selected clusters. Table-wise Bonferroni corrected $\alpha = 0.0013$.)

label	variable	3	8	pval	test
original_shape_VoxelVolume	Mean (std)	0.4 (0.1)	0.9 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHL_glcm_ClusterProminence	Mean (std)	0.2 (0.1)	0.1 (0.04)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_glszm_LargeAreaLowGrayLevelEmphasis	Mean (std)	0.05 (0.05)	0.5 (0.2)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHL_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.3 (0.1)	0.8 (0.1)	<0.0001	(Two Sample t-test)
wavelet.HHH_glszm_GrayLevelNonUniformity	Mean (std)	0.02 (0.03)	0.1 (0.1)	<0.0001	(Wilcoxon rank sum test)
wavelet.HLL_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.4 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.LHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.1 (0.05)	0.01 (0.009)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHH_glcm_ClusterProminence	Mean (std)	0.1 (0.05)	0.02 (0.01)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LLH_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.4 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.HHH_glszm_LargeAreaHighGrayLevelEmphasis	Mean (std)	0.4 (0.1)	0.8 (0.1)	<0.0001	(Welch Two Sample t-test)
wavelet.LLL_glcm_ClusterProminence	Mean (std)	0.7 (0.1)	0.5 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHL_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.1 (0.04)	0.03 (0.02)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHL_gldm_SmallDependenceLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.1 (0.04)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHH_glcm_Imc1	Mean (std)	0.8 (0.1)	0.5 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.2.0.mm.3D_glrlm_LongRunHighGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.5 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.LHH_glszm_GrayLevelNonUniformity	Mean (std)	0.1 (0.1)	0.4 (0.2)	<0.0001	(Welch Two Sample t-test)
wavelet.LHH_glcm_Imc1	Mean (std)	0.6 (0.2)	0.8 (0.1)	<0.0001	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_firstorder_Kurtosis	Mean (std)	0.05 (0.03)	0.2 (0.2)	<0.0001	(Wilcoxon rank sum exact test)
wavelet.HHH_glrlm_GrayLevelVariance	Mean (std)	0.1 (0.1)	0.04 (0.01)	0.0001	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_firstorder_90Percentile	Mean (std)	0.8 (0.1)	0.9 (0.02)	0.0002	(Wilcoxon rank sum exact test)
clinical.T.Stage	1	15 (100.00%)	0 (0%)	0.0003	(Fisher's Exact Test for Count Data)
clinical.T.Stage	2	12 (44.44%)	15 (55.56%)	0.0003	(Fisher's Exact Test for Count Data)
clinical.T.Stage	3	3 (42.86%)	4 (57.14%)	0.0003	(Fisher's Exact Test for Count Data)
clinical.T.Stage	4	7 (41.18%)	10 (58.82%)	0.0003	(Fisher's Exact Test for Count Data)
wavelet.LHL_gldm_DependenceVariance	Mean (std)	0.2 (0.2)	0.3 (0.2)	0.0004	(Wilcoxon rank sum exact test)
wavelet.HLH_glcm_DifferenceAverage	Mean (std)	0.4 (0.2)	0.3 (0.03)	0.0008	(Wilcoxon rank sum exact test)
wavelet.HHH_firstorder_RootMeanSquared	Mean (std)	0.7 (0.1)	0.8 (0.02)	0.0008	(Wilcoxon rank sum exact test)
wavelet.HHL_gldm_SmallDependenceHighGrayLevelEmphasis	Mean (std)	0.1 (0.05)	0.1 (0.03)	0.0042	(Welch Two Sample t-test)
log.sigma.5.0.mm.3D_glszm_SizeZoneNonUniformityNormalized	Mean (std)	0.5 (0.2)	0.4 (0.1)	0.0045	(Two Sample t-test)
wavelet.HLH_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.2 (0.1)	0.1 (0.1)	0.0095	(Wilcoxon rank sum exact test)
gender	female	16 (76.19%)	5 (23.81%)	0.0244	(Pearson's Chi-squared test)
gender	male	21 (46.67%)	24 (53.33%)	0.0244	(Pearson's Chi-squared test)
age	Mean (std)	0.6 (0.2)	0.7 (0.1)	0.0747	(Two Sample t-test)
Manufacturer	CMS Inc.	10 (41.67%)	14 (58.33%)	0.0749	(Pearson's Chi-squared test)
Manufacturer	SIEMENS	27 (64.29%)	15 (35.71%)	0.0749	(Pearson's Chi-squared test)
wavelet.HHH_glszm_GrayLevelNonUniformityNormalized	Mean (std)	0.2 (0.2)	0.3 (0.2)	0.3322	(Wilcoxon rank sum test)
Overall.Stage	I	11 (61.11%)	7 (38.89%)	0.3410	(Fisher's Exact Test for Count Data)
Overall.Stage	II	1 (25.00%)	3 (75.00%)	0.3410	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIa	10 (71.43%)	4 (28.57%)	0.3410	(Fisher's Exact Test for Count Data)
Overall.Stage	IIIb	15 (50.00%)	15 (50.00%)	0.3410	(Fisher's Exact Test for Count Data)
original_glcm_Imc1	Mean (std)	0.6 (0.1)	0.6 (0.2)	0.4332	(Wilcoxon rank sum exact test)
log.sigma.2.0.mm.3D_glcm_Autocorrelation	Mean (std)	0.2 (0.1)	0.2 (0.1)	0.4720	(Wilcoxon rank sum exact test)
Histology	adenocarcinoma	5 (55.56%)	4 (44.44%)	0.6434	(Fisher's Exact Test for Count Data)
Histology	large cell	16 (64.00%)	9 (36.00%)	0.6434	(Fisher's Exact Test for Count Data)
Histology	nos	10 (55.56%)	8 (44.44%)	0.6434	(Fisher's Exact Test for Count Data)
Histology	squamous cell carcinoma	6 (42.86%)	8 (57.14%)	0.6434	(Fisher's Exact Test for Count Data)
Clinical.N.Stage	0	15 (50.00%)	15 (50.00%)	0.6593	(Pearson's Chi-squared test)
Clinical.N.Stage	1_2	13 (61.90%)	8 (38.10%)	0.6593	(Pearson's Chi-squared test)
Clinical.N.Stage	3	9 (60.00%)	6 (40.00%)	0.6593	(Pearson's Chi-squared test)
wavelet.LHH_firstorder_Skewness	Mean (std)	0.5 (0.1)	0.5 (0.1)	0.9591	(Wilcoxon rank sum exact test)
log.sigma.5.0.mm.3D_glszm_SmallAreaLowGrayLevelEmphasis	Mean (std)	0.04 (0.03)	0.03 (0.01)	0.9898	(Wilcoxon rank sum exact test)

Table 5.21: Table comparing distribution for different variables between clusters 7 and 9, from the model without variable selection, those showing greatest difference in median survival. Both variables included as input in the model or not are evaluated according to its distribution amongst selected clusters. Table-wise Bonferroni corrected $\alpha = 0.0013$.)

Chapter 6

Discussion

This dataset is of main relevance in radiomic research for its unique size and availability of manual segmentations. It was previously used in several publications exploring radiomic features relevance in NSCLC patients. Most of previous work using this dataset is mainly centered on generating predictive models or identifying radiomic signatures, but, no detailed analysis of clinical and main scanner characteristics, to better understand this dataset and potential biases, seems to be available (Aerts et al., 2014; Wu et al., 2016; Shi et al.,2019). During the present analysis forty-two patients presented missing values for Histology variable, one of the main target variables in this data set and related to main inclusion criteria, the diagnosis of histologically confirmed NSCLC. NOS (not otherwise specified) class is included as a possible value for Histology variable, so it is not explicitly clear if patients with missing values had diagnosis of NSCLC confirmed or not. This is the main reason why exclusion of these patients was decided before further analysis of this dataset. Isolated erroneous entries were identified for clinical T, N and M stage variables as well as for manufacturer model variables, missing values were assigned to these entries in the present work, and treated as other missing values with the selected imputation technique. When testing for independence between different clinical and main scanner variables, I found a statistically significant association between histology and age and between histology and overall stage. This lack of independence may add a bias when interpreting the relationship between radiomic features related to one variable or the other on its own, without taking the others into account, as conclusions assigned to one variable could correspond to the other and vice versa. It is important to take into account this relationship and also evaluate results for both variables when assessing one or the other. In addition, there was a significant difference between survival probability curves between patients with CT studies performed in different scanner models used. This could include a technical source of bias that should be compensated or accomplished when performing further analysis. In the present work I chose to eliminate all radiomic features possibly related to scanner manufacturer trying to retain only stable features for further analysis. I chose not to correct α value when testing for this relationship risking to lose some relevant variable in the benefit of compensating this source of bias. We should take into account though that some relevant features related to target variables were probably lost within this first step of variable selection.

When performing clustering analysis original shape VoxelVolume dominated both models

definition along with log.sigma.5.0.mm.3D glzm LargeAreaLowGrayLevelEmphasis and wavelet LHL glzm LargeAreaHighGrayLevelEmphasis. As expected, given its direct relationship with size, Clinical T stage showed a significant association with most of the partitions compared. Both models showed significant association between partitions generated and survival curves, significance of this association was greater for the model with variable selection which finally included only radiomic variables. Clusters with significant lower median survival were also related to higher Clinical T stages, greater mean values of original_shape_VoxelVolume (volume of the ROI), log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis (Texture Gray Level Size Zone Matrix Features that measure the proportion of the joint distribution of larger size zones with lower or higher gray-level values in the ROI respectively) and lower mean wavelet.HHl_glcm_ClusterProminence (Texture measure of the skewness and asymmetry of the Gray Level Co-occurrence Matrix). The opposite was true for clusters with greatest median survival. There was no significant relationship between manufacturer model and clusters compared for survival analysis. Along with previous selection of stable features regarding manufacturer, this favors a more transparent interpretation of the results. Overall stage was associated with clusters with similar characteristics to these, with higher stages predominating in the same clusters with higher T category, but was not significantly associated with clusters generated by minimum and maximum median survival.

Histology categories showed only weak association with selected clusters regarding the model without variable selection. Within cluster with higher proportion of squamous cell carcinoma, showed greater original_shape_VoxelVolume, higher mean wavelet.LHL_gldm_DependenceVariance (Texture measure of the variance in dependence size in the image, dependency being defined as the number of connected pixels that are similar to a center pixel) and lower mean wavelet.HHH_glcm_Imc1 (a measure that quantifies the complexity of the texture). The opposite was true for clusters with higher proportions of adenocarcinoma and nos histotypes. Large cell carcinoma was equally distributed within clusters selected for histology comparisons.

Some limitations in the analysis of this data should be taken into account. Source of segmentation masks used to extract radiomic features was the previous manual definition of the region of interest by a single operator and, though this eliminates the risk of inter-operator variability, we do not have the elements to estimate the magnitude of intra-operator variability. Regarding α value correction, depending on the magnitud to which we want to be conservative, correction could be applied for the total of tests done along the whole analysis, risking to lose some relevant information in expense of increasing false negatives. I chose to give the reader the corrected α value table-wise to provide additional information for the interpretation of each table without leaving aside information that I thought could be relevant, but this should be taken into account when reading the results. When fitting model based clustering, VarselLCM assumes variable independence when applying its wrapper model selection method, so it may tend to retain more variables than necessary when using as input a dataset like the one presented where variables are correlated.

6.1 Conclusion

In conclusion, potential sources of bias given by relationship between different variables of interest and technical sources should be taken into account when analyzing this data set. Aside from original_shape_VoxelVolume feature, texture features applied to images with LoG and wavelet filters were found most significantly associated with different clinical characteristics in the present analysis. When performing clustering analysis, partitions were mainly related to survival finding some interesting association within minimum and maximum medium survival defined clusters. original_shape_VoxelVolume, log.sigma.5.0.mm.3D_glzm_LargeAreaLowGrayLevelEmphasis and wavelet_LHL_glzm_LargeAreaHighGrayLevelEmphasis were the variables with most discriminative power as calculated when generating the model.

Radiomics can quantify tumor phenotypic characteristics non-invasively and could potentially contribute with objective elements to support these patients' diagnosis, management and prognosis in routine clinical practice. This work aims to describe associations and not causality, being descriptive, it establishes the base for future work.

The main lesson learned from this work was the relevance of doing a good exploratory analysis of the data before further deepening the analysis. With the current work I learned how to extract radiomic features from medical images and I was able to apply these techniques myself to build a rich dataset. I was mainly interested in understanding and being able to apply model based clustering and different visualization techniques for multivariate data. I was able to learn an apply new algorithms for missing values imputation, multivariate outlier detection and model based clustering. Regarding initial plans and goal, extraction of radiomic features itself to be able to build the dataset took me much more time than expected, leaving me less time than what I initially planned for the second part of the work.

6.2 Future work

To continue with this work I would like to formally compare different feature selection methods and clustering techniques to analyze these data, other than the ones used in the present work.

6.3 Follow-up of planning

Regarding initial plans and goal, extraction of radiomic features itself to be able to build the dataset took me much more time than expected, leaving me less time than what I initially planned to analyze the data itself so I had to focus on a specific analysis. Though I explored other techniques while performing the present work I would like to formally compare more techniques for feature selection and for clustering of the data.

Chapter 7

Glosary

- NSCLC: Non-small cell lung cancer.
- UICC: The Union for International Cancer Control
- NOS: Not otherwise specified
- TNM: Tumor Node Metastasis
- CT: Computed tomography
- DICOM: Digital Imaging and Communications in Medicine
- GLCM: Gray Level Cooccurence Matrix
- GLRLM: Gray Level Run Length Matrix
- GLSZM: Gray Level Size Zone Matrix
- GLDM: Gray Level Dependece Matrix
- LoG: Laplacian of Gaussian
- H: High pass filter
- L: Low pass filter
- MRMR: Maximum Relevance Minimum Redundancy
- PCA: Principal Component analysis
- CA: Correspondence analysis

References

- Jani C., Marshall D.C., Singh H., Goodall R., Shalhoub J., Omari O., et al. (2021) Lung cancer mortality in Europe and the USA between 2000 and 2017: an observational analysis. ERJ Open Research. 7: 00311-2021; https://doi.org/10.1183/23120541.00311-2021.
- Gridelli, C., Rossi, A., Carbone, D., Guarize j., Karachaliou N., Mok T., et al. (2015) Non-small-cell lung cancer. Nat Rev Dis Primers 1, 15009. https://doi.org/10.1038/nrdp.2015.9
- Purandare, N. C., & Rangarajan, V. (2015). Imaging of lung cancer: Implications on staging and management. The Indian journal of radiology & imaging, 25(2), 109–120. https://doi.org/10.4103/0971-3026.155831
- Scrivener, M., de Jong, E. E., van Timmeren, J. E., Pieters, T., Ghaye, B., & Geets, X. (2016). Radiomics applied to lung cancer: a review. Transl Cancer Res, 5(4), 398-409.
- Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Cavalho, S., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications. Nature Publishing Group. http://doi.org/10.1038/ncomms5006
- Junior, J. R. F., Koenigkam-Santos, M., Cipriano, F. E. G., Fabro, A. T., & de Azevedo-Marques, P. M. (2018). Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. Computer methods and programs in biomedicine, 159, 23-30.
- Luna, J. M., Barsky, A. R., Shinohara, R. T., Roshkovan, L., Hershman, M., Dreyfuss, A. D., ... & Kontos, D. (2022). Radiomic Phenotypes for Improving Early Prediction of Survival in Stage III Non-Small Cell Lung Cancer Adenocarcinoma after Chemoradiation. Cancers, 14(3), 700.
- Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., & Bellomi, M. (2018). Radiomics: the facts and the challenges of image analysis. European radiology experimental, 2(1), 1-8.
- Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., ... & Löck, S. (2020). The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology, 295(2), 328-338.

- van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. Insights into imaging, 11(1), 1-16.
- Lee, S. H., Cho, H. H., Lee, H. Y., & Park, H. (2019). Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. Cancer Imaging, 19(1), 1-12.
- Liu, H., Li, H., Habes, M., Li, Y., Boimel, P., Janopaul-Naylor, J., ... & Fan, Y. (2020). Robust collaborative clustering of subjects and radiomic features for cancer prognosis. IEEE Transactions on Biomedical Engineering, 67(10), 2735-2744.
- Yousefi, B., Jahani, N., LaRiviere, M. J., Cohen, E., Hsieh, M. K., Luna, J. M., ... & Kontos, D. (2019, March). Correlative hierarchical clustering-based low-rank dimensionality reduction of radiomics-driven phenotype in non-small cell lung cancer. In Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications (Vol. 10954, pp. 278-285). SPIE.
- Aerts, H. J. W. L., Wee, L., Rios Velazquez, E., Leijenaar, R. T. H., Parmar, C., Grossmann, P., et al. (2019). Data From NSCLC-Radiomics. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI
- Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. (2013) The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, pp 1045-1057.
- R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
- Barta, J. A., Powell, C. A., & Wisnivesky, J. P. (2019). Global epidemiology of lung cancer. Annals of global health, 85(1).
- Brierley, J. D., Gospodarowicz, M. K., & Wittekind, C. (Eds.). (2017). TNM classification of malignant tumours. John Wiley & Sons.
- Postmus, P. E., Kerr, K. M., Oudkerk, M., Senan, S., Waller, D. A., Vansteenkiste, J., ... & Peters, S. (2017). Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology, 28, iv1-iv21.
- Righi, L., Vavalà, T., Rapa, I., Vatrano, S., Giorcelli, J., Rossi, G., ... & Papotti, M. (2014). Impact of non–small-cell lung cancer-not otherwise specified immunophenotyping on treatment outcome. Journal of Thoracic Oncology, 9(10), 1540-1546
- Detterbeck, F. C., Boffa, D. J., Kim, A. W., & Tanoue, L. T. (2017). The eighth edition lung cancer stage classification. Chest, 151(1), 193-203.

- Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., & Cook, G. (2020). Introduction to radiomics. Journal of Nuclear Medicine, 61(4), 488-495.
- Papanikolaou, N., Matos, C., & Koh, D. M. (2020). How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging, 20(1), 1-10.
- Depeursinge, A., Andrearczyk, V., Whybra, P., van Griethuysen, J., Müller, H., Schaer, R., ... & Zwanenburg, A. (2020). Standardised convolutional filtering for radiomics. arXiv preprint arXiv:2006.05470.
- Fournier, L., Costaridou, L., Bidaut, L., Michoux, N., Lecouvet, F. E., de Geus-Oei, L. F., ... & European Society of Radiology. (2021). Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers. European radiology, 31(8), 6001-6012.
- Fornacon-Wood, I., Mistry, H., Ackermann, C. J., Blackhall, F., McPartlin, A., Faivre-Finn, C., ... & O'Connor, J. P. (2020). Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. European radiology, 30(11), 6241-6250.
- Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. Cancer research, 77(21), e104-e107.
- Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer Science & Business Media.
- Landau, S., Leese, M., Stahl, D., & Everitt, B. S. (2011). Cluster analysis. John Wiley & Sons.
- Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. Statistics Surveys, 12, 18-65.
- Fournier, L., Costaridou, L., Bidaut, L., Michoux, N., Lecouvet, F. E., de Geus-Oei, L. F., ... & European Society of Radiology. (2021). Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers. European radiology, 31(8), 6001-6012.
- Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering. Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra.
- Haga, A., Takahashi, W., Aoki, S., Nawa, K., Yamashita, H., Abe, O., & Nakagawa, K. (2019). Standardization of imaging features for radiomics analysis. The Journal of Medical Investigation, 66(1.2), 35-37.

- Sharp, G. C., Li, R., Wolfgang, J., Chen, G., Peroni, M., Spadea, M. F., et al (2010). Plastimatch: an open source software suite for radiotherapy image processing. In Proceedings of the XVI'th International Conference on the use of Computers in Radiotherapy (ICCR), Amsterdam, Netherlands.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.
- Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2021). naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.6.1. https://CRAN.R-project.org/package=naniar
- Boxuan Cui (2020). DataExplorer: Automate Data Exploration and Treatment. R package version 0.8.2. https://CRAN.R-project.org/package=DataExplorer
- C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.
- Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork
- Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2022). skimr: Compact and Flexible Summaries of Data. R package version 2.1.4. https://CRAN.R-project.org/package=skimr
- Choonghyun Ryu (2022). dlookr: Tools for Data Diagnosis, Exploration, Transformation. R package version 0.5.6. https://CRAN.R-project.org/package=dlookr
- Dan Chaltiel (2022). crosstable: Crosstables for Descriptive Analyses. R package version 0.4.1. https://CRAN.R-project.org/package=crosstable
- Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra
- David Gohel (2022). flextable: Functions for Tabular Reporting. R package version 0.7.0. https://CRAN.R-project.org/package=flextable
- Daniel J. Stekhoven (2022). missForest: Nonparametric Missing Value Imputation using Random Forest.R package version 1.5.
- Stekhoven, D.J. and Buehlmann, P. (2012), 'MissForest - nonparametric missing value imputation for mixed-type data', Bioinformatics, 28(1) 2012, 112-118, doi: 10.1093/bioinformatics/btr597

- Chris Fraley (2022). HDoutliers: Leland Wilkinson's Algorithm for Detecting Multidimensional Outliers. R package version 1.0.4. https://CRAN.R-project.org/package=HDoutliers
- Wilkinson, L. (2017). Visualizing big data outliers through distributed aggregation. IEEE transactions on visualization and computer graphics, 24(1), 256-266.
- Max Kuhn (2022). caret: Classification and Regression Training. R package version 6.0-92. https://CRAN.R-project.org/package=caret
- Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.
- Miron B. Kursa (2021). Praznik: High performance information-based feature selection. SoftwareX, 16, 100819. URL https://doi.org/10.1016/j.softx.2021.100819
- N De Jay, S Papillon-Cavanagh, C Olsen, G Bontempi and B Haibe-Kains mRMRe: an R package for parallelized mRMR ensemble feature selection Submitted (2012).
- Marbac, M., & Sedki, M. (2019). VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values. Bioinformatics, 35(7), 1255-1257.
- Marbac, M. and Patin, E. and Sedki, M. (2020). Variable selection for mixed data clustering: Application in human population genomics. Journal of Classification. 37:124–142. https://doi.org/10.1007/s00357-018-9301-y
- Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2022). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.3 — For new features, see the 'Changelog' file (in the package source), https://CRAN.R-project.org/package=cluster
- Nenadic, O., Greenacre, M. (2007) Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software 20(3):1-13.
- Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra
- Therneau T (2022). A Package for Survival Analysis in R. R package version 3.3-1, https://CRAN.R-project.org/package=survival.
- Alboukadel Kassambara, Marcin Kosinski and Przemyslaw Biecek (2021). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.9. https://CRAN.R-project.org/package=survminer
- Wu, W., Parmar, C., Grossmann, P., Quackenbush, J., Lambin, P., Bussink, J., ... & Aerts, H. J. (2016). Exploratory study to identify radiomics classifiers for lung cancer histology. Frontiers in oncology, 6, 71.

- Shi, Z., Zhovannik, I., Traverso, A., Dankers, F. J., Deist, T. M., Kalendralis, P., ... & Wee, L. (2019). Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Scientific data, 6(1), 1-8.